Essays on Collective Intelligence and the Future of Work

MASSACHUSETTS INSTITUTE
OF TECHNOLOGY
JUL 25 2019
LIBRARIES
ARCHIVES

by

Erik P. Duhaime

A.B., Economics and Human Biology
Brown University, 2010

M.Phil., Human Evolution
University of Cambridge, 2011

Submitted to the Sloan School of Management
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY IN MANAGEMENT

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2019

©2019 Massachusetts Institute of Technology. All rights reserved.

Signature of Author: Signature redacted

Erik P. Duhaime
Department of Management
May 2, 2019

Certified by: Signature redacted

Thomas W. Malone
Patrick J. McGovern Professor of Management
Thesis Supervisor

Accepted by: Signature redacted

Ezra Zuckerman Sivan
Alvin J. Siteman (1948) Professor of Entrepreneurship and Strategy
Deputy Dean
Faculty Chair, MIT Sloan PhD Program

MITLibraries

77 Massachusetts Avenue
Cambridge, MA 02139
http://libraries.mit.edu/ask

DISCLAIMER NOTICE

Due to the condition of the original material, there are unavoidable flaws in this reproduction. We have made every effort possible to provide you with the best copy available.

Thank you.

The images contained in this document are of the best quality available.

Essays on Collective Intelligence and the Future of Work

by

Erik P. Duhaime

Submitted to the Sloan School of Management on May 2, 2019 in partial fulfillment of the requirements for the degree of Doctor of Philosophy

Abstract:
This dissertation considers how information technologies have enabled new ways of organizing work. The first essay investigates how artificial intelligence (AI) systems will impact the field of medical diagnostics. While previous research has shown that AI can diagnose skin cancer as accurately as professional dermatologists, I explore what happens when AI is combined with—rather than compared to—human intelligence. Using a dataset of the diagnoses of 1 state-of-the-art AI system and 21 board-certified dermatologists on 371 biopsy-proven cases of skin lesions, I find that averaging the opinion of an individual dermatologist with AI often does not lead to higher accuracy than AI alone. However, combining AI with the average opinion from groups of dermatologists leads to higher performance than individuals alone, AI alone, and groups alone. This suggests that in many cases artificial intelligence will not simply replace jobs, but rather, will transform how work is organized.

The second essay (coauthored with B. Bond, Q. Yang, P. de Boer, and T.W. Malone) considers how crowdsourcing can be used to find innovative solutions to complex problems. In past research, recursive incentive schemes have shown promise for conducting social search by motivating people to use their weak ties to find distant targets, such as specific people or even weather balloons placed at undisclosed locations. We report on a case study of a similar recursive incentive scheme for finding innovative ideas. Specifically, we implemented a competition to reward individual(s) who helped refer Grand Prize winner(s) in MIT's Climate CoLab, an open innovation platform for addressing global climate change. Using data on over 78,000 CoLab members and over 36,000 people from over 100 countries who engaged with the referral contest, we find that people who are referred using this method are more likely than others both to submit proposals and to submit high quality proposals. Furthermore, we find suggestive evidence that among the contributors referred via the contest, those who had more than one degree of separation from a pre-existing CoLab member were more likely to submit high quality proposals. Thus, the results from this case study are consistent with the theory that people from distant networks are more likely to provide innovative solutions to complex problems.

The third essay (coauthored with Z. Woessner) considers how newly enabled organizational designs are changing the social norms and expectations of workers. Specifically, we investigate the social norm of tipping and propose that work in the "gig economy" is associated with a breakdown of tipping norms in part because of workers' increased autonomy in terms of deciding when and whether to work. We present four studies to support our hypothesis: a survey vignette experiment with workers on Amazon Mechanical Turk (Study 1), an analysis of New York City taxi data (Study 2), a field experiment with restaurant employee food delivery drivers (Study 3), and a field experiment with gig-worker food delivery drivers (Study 4). In Studies 1 and 2, we find that consumers are less likely to tip when workers have autonomy in deciding whether to complete a task. In Study 3, we find that restaurant delivery employees notice upfront tips (or lack thereof) and alter their service as a result. In contrast, in Study 4, we find that gig-workers who agree to complete a delivery for a fixed amount that includes an upfront tip (or lack thereof) are not responsive to tips. Together, these findings suggest that the gig economy has not only transformed employee-employer relationships, but has also altered the norms and expectations of consumers and workers.

Thesis Supervisor: Thomas W. Malone
Title: Patrick J. McGovern Professor of Management

Acknowledgements

It certainly takes a village.

I am extremely grateful to the many people who helped make this dissertation possible. First and foremost, I'd like to thank my advisor Tom Malone for his unwavering support, guidance and inspiration. When I first came to MIT I considered myself an expert in the evolutionary biology of cooperation, but I had not thought much about artificial intelligence or how information technology could enable new ways of cooperating. There is no doubt that Tom's mentorship has had an incredible influence on where I am today, and what I will accomplish in the future.

I am also extremely grateful to my committee members, John Carroll and David Rand, as well as the other Sloan faculty who I have learned from and who have supported me during my time at MIT. Jared Curhan, Evan Apfelbaum, and Ray Reagans have been especially kind, thoughtful, and helpful. I'm also grateful to Liz McFall, Richard Hill, Hillary Ross, and Davin Lee Schnappauf. A number of fellow PhD students have also been there with me through thick and thin: Emily Truelove, Jenna Myers, Heather Yang, Minjae Kim, Sam Zyontz, Nathan Wilmers, Maxim Massenkoff, Jason Nemirow, Rebecca Grunberg, Daniel Kim, Daniel Rock, Brittany Bond, Vanessa Conzon, and the rest of my initial BPS cohort--Taylor Moulton, Jorge Guzman, Christine Riordan, Hyejun Kim.

Of course, I would also not be here today were it not for the support of Olivia, my parents, and my friends. Thank you.

This page intentionally left blank

Table of Contents

This page intentionally left blank

ESSAY 1

Human-Computer Groups Outperform Artificial Intelligence at Diagnosing Skin Cancer

Erik P. Duhaime

Abstract:

While previous research has shown that AI can diagnose skin cancer as accurately as professional dermatologists, I explore what happens when AI is combined with—rather than compared to—human intelligence. Using a dataset of the diagnoses of 1 state-of-the-art AI system and 21 board-certified dermatologists on 371 biopsy-proven cases of skin lesions, I find that averaging the opinion of an individual dermatologist with AI often does not lead to higher accuracy than AI alone. However, combining AI with the average opinion from groups of dermatologists leads to higher performance than individuals alone, AI alone, and groups alone. This suggests that in many cases artificial intelligence will not simply replace jobs, but rather, will transform how work is organized.

Human-Computer Groups Outperform Artificial Intelligence at Diagnosing Skin Cancer

Computers can now diagnose diseases such as lung and colorectal cancer (Somashekhar et al., 2017), detect pneumonia in chest x-rays at a level exceeding practicing radiologists (Rajpurkar et al., 2017), predict heart attacks and strokes more accurately than doctor's standard methods (Weng, Reps, Kai, Garibaldi, & Qureshi, 2017), diagnose congenital cataracts as accurately as human ophthalmologists (Long et al., 2017), identify signs of autism spectrum disorder in brain scans of infants (Hazlett et al., 2017), and identify breast cancer as accurately as radiologists (Becker et al., 2017). As a result of this rapid progress, a growing chorus of academics, thought leaders, and politicians warn that artificial intelligence will cause unprecedented job loss, even among highly trained knowledge workers such as doctors.

At the same time, advances in information technology have made it easier than ever to harness the collective intelligence of groups of people throughout the world, thereby enabling entirely new ways of organizing work (Malone, 2004; Surowiecki, 2005). A robust scientific literature on "the wisdom of crowds" shows that combining the decisions of multiple people is oftentimes better than the decisions of the best person in the group, and a simple average has proven to be an effective aggregation method in a wide range of situations (Armstrong, 2001; Clemen & Winkler, 1986; Galton, 1907). Recent research has successfully applied this concept to medical diagnostics by demonstrating that averaging the opinions of multiple doctors can detect breast and skin cancer more accurately than any individual doctor (Kurvers et al., 2016; Wolf, Krause, Carney, Bogart, & Kurvers, 2015). However, it remains unknown whether—and when—mixed human-computer groups will outperform either approach alone.

I examine this question by considering the case of diagnosing skin cancer, the most common malignancy in the United States (Rogers, 2015). In one of the most impressive recent demonstrations of the potential for artificial intelligence to revolutionize medical diagnostics, a team of researchers at Stanford University trained a convolutional neural net (CNN) using a dataset of 129,450 clinical images and then tested its performance against board-certified dermatologists on biopsy-proven clinical images for two important clinical use cases: 1) identifying keratinocyte carcinomas, the most common cancer in the United States, and 2) identifying malignant melanomas, the deadliest skin cancer (Esteva et al., 2017). In their study, at least 21 dermatologists analyzed each of 376 total images: 135 photographic images for carcinomas (65 keratinocyte carcinomas, 70 benign seborrheic keratoses), 130 photographic images for melanoma (33 malignant melanomas, 97 benign nevi), and 111 dermoscopic images for melanoma (71 malignant, 40 benign). Figure 1 demonstrates the difficulty of correctly classifying malignant and benign skin lesions.

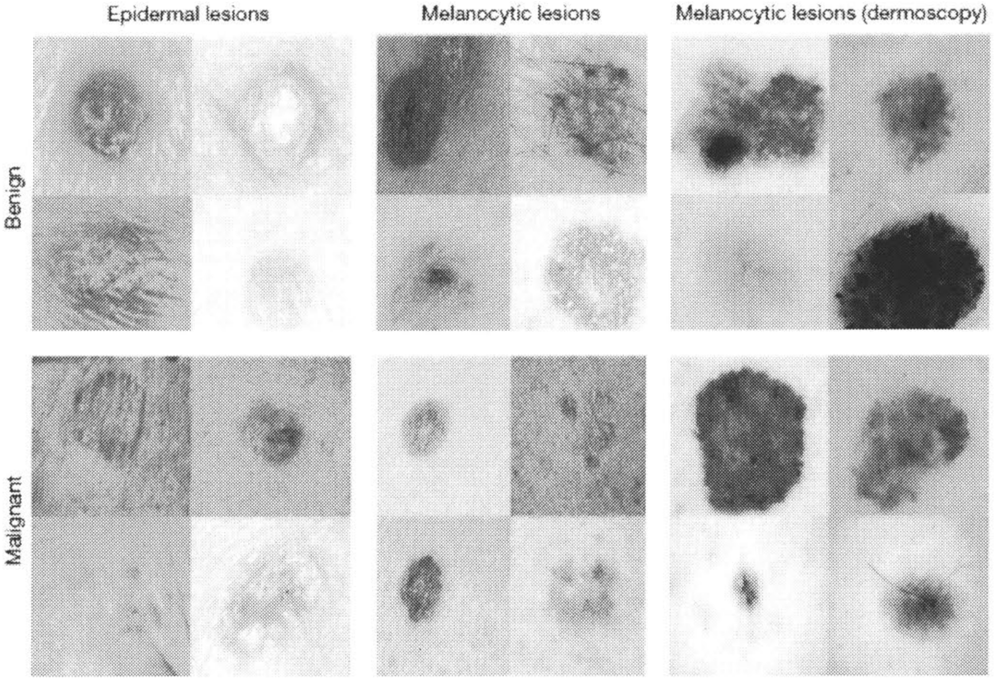
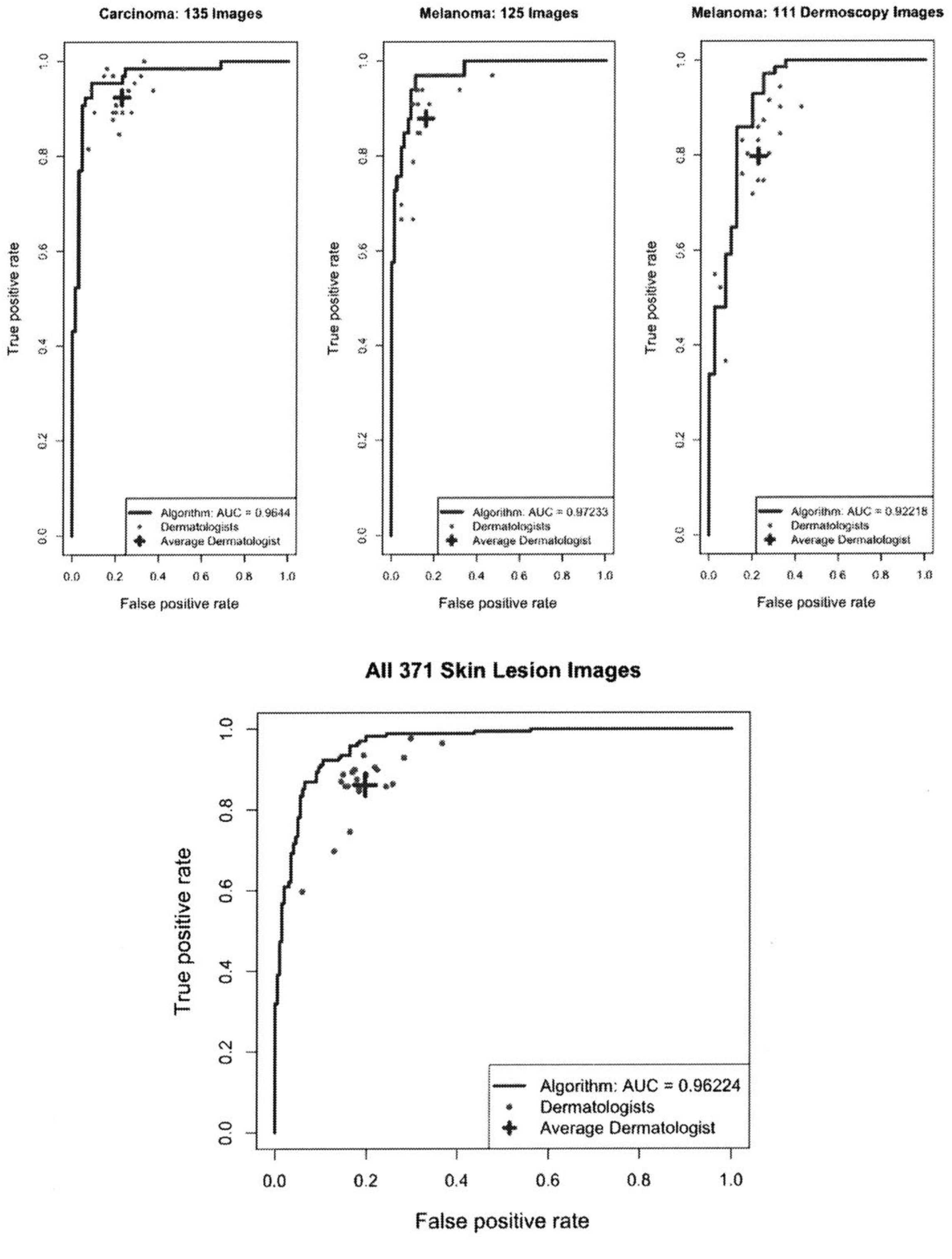


Figure 1. Malignant and benign example images of skin lesions. These test images highlight the difficulty of malignant versus benign discernment for the three medically critical classification tasks (image from Esteva et al., 2017).

While 31 board-certified dermatologists analyzed at least some of the 376 images and at least 21 of them analyzed each image, only 1 dermatologist analyzed all 130 of the photographic images for melanoma. I remove 5 photographic images from the melanoma test as well as the diagnoses of dermatologists who did not analyze all of the remaining 371 images. This results in a slightly smaller but much cleaner dataset of 21 dermatologists who all indicate whether or not they think a case is cancerous (0 or 1) on the same 371 cases. While the dermatologists' diagnoses are binary (i.e., 0 or 1), the CNN outputs a probability *P* per image, which can be transformed into a binary classification by fixing a threshold probability *t*. By increasing *t* from 0 to 1, I am able to draw an ROC (Receiver Operating Characteristic) curve and calculate the AUC (Area

Under the Curve) to evaluate the performance of the CNN. ROC curves for each of the three different image types are depicted in Figure 2A. Figure 2A also shows each dermatologist's diagnostic accuracy, depicted as a point based on the true positive rate (i.e., the sensitivity) and the false positive rate (i.e., 1 minus the specificity). Figure 2B depicts the same information with all case types combined into one larger sample. As can be seen in Figures 2A and 2B, while some individual dermatologists outperform the CNN at some of the three image analysis tasks, the algorithm achieves accuracy at or above the level of each of the dermatologists on the combined dataset.



Figures 2A & 2B. A comparison of diagnostic accuracy of 21 board-certified dermatologists (red dots) and one state-of-the-art algorithm (blue line) across three diagnostic tasks (Figure 2A, above) and on the three tasks combined (Figure 2B, below). The green cross represents the average dermatologist accuracy.

To compare the accuracy of the CNN to that of groups of dermatologists, I randomly sample *n* of the 21 dermatologists to create a group, average their opinions to form a probability, *P*, for each of the 371 cases, plot an ROC curve, and then calculate the AUC. I repeat the process 1,000 times for each group size and plot the average AUC achieved in Figure 3, below. Consistent with past research, small groups of as few as 2-6 people are sufficient to observe substantial improvements over individual decision makers (Ariely et al., 2000). As group size increases, the average group accuracy approaches that of the CNN. In fact, according to a DeLong's test for two correlated ROC curves (DeLong, DeLong, & Clarke-Pearson, 1988), the group of all 21 dermatologists is not significantly less accurate than the CNN alone ($p = .62$).

Next, rather than compare the CNN to groups of dermatologists, I create human-computer "centaur" groups by combining the two using the 50/50 model (Blattberg & Hoch, 1990). As shown in Figure 3, individuals combined with the CNN achieve lower average accuracy than the CNN alone ($t(20) = -2.84$, $p = .01$). In contrast, groups of dermatologists combined with the CNN tend to achieve higher accuracy than both the human groups and the CNN alone. For instance, only 7 of the 21 individual dermatologist centaurs but 76.0% of the 1,000 simulated two person centaur groups, 96.8% of the 1,000 simulated three person centaur groups, 99.99% of the 1,000 simulated four person centaur groups, and 100% of the centaur groups with more than 4 dermatologist opinions achieve a higher AUC than the CNN alone. Furthermore, according to a DeLong's test for two correlated ROC curves (DeLong et al., 1988), the 21-dermatologist centaur model achieves significantly greater AUC than both the group of 21 dermatologists alone ($p = .002$) and the CNN alone ($p = .01$). The difference is substantial: the 21-dermatologist centaur model can achieve a sensitivity of over 99% with a false positive rate of only 21%, less than half the false positive rate required by the dermatologist group alone (54%) and the CNN alone (44%).

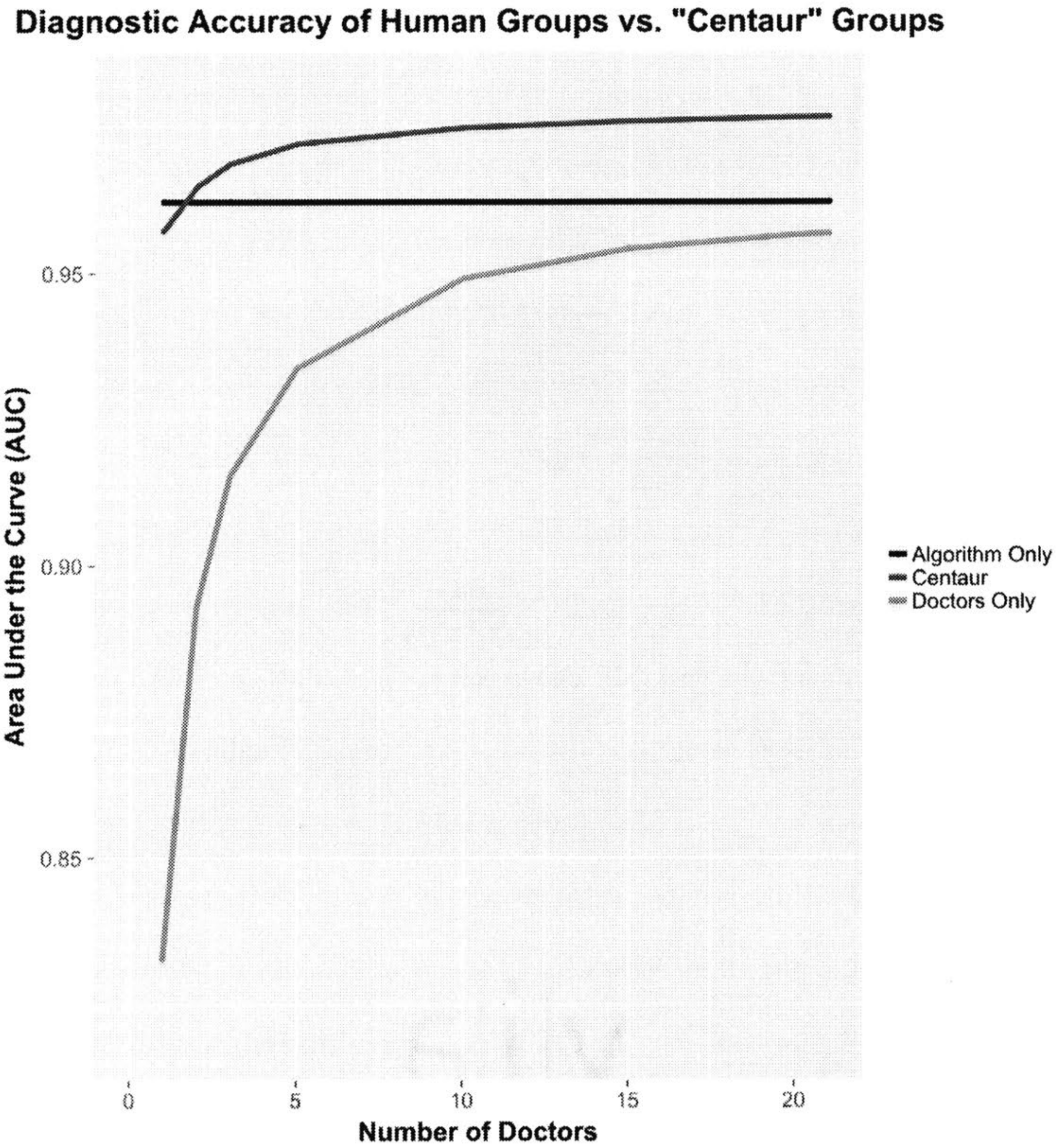


Figure 3. Increases in diagnostic accuracy as a function of number of dermatologist's opinions averaged together, both with (red line) and without (blue line) being combined with the CNN.

It is notable that a single doctor's opinion averaged with the CNN achieves lower accuracy than the CNN alone. As shown in Figure 4, when an individual doctor's opinion is averaged with the CNN, the corresponding ROC curve always passes through the point indicating that doctor's performance on their own, which is below the ROC curve of the CNN. The reason for this is that while the CNN provides a probabilistic diagnosis, the doctors' diagnoses are binary. Therefore, for each ROC curve there is a point—specifically, the point at which the individual doctor is calibrated to trade off false positives vs. false negatives—where the threshold value used to create the ROC curve is .5, and at that point the CNN's diagnosis will have no impact because—when averaged—it will be drowned out by the doctor's diagnosis of either a 0 or 1. In other words, there is a tension over decision authority that is resolved in favor of the individual doctor's opinion, rather than the more accurate CNN.

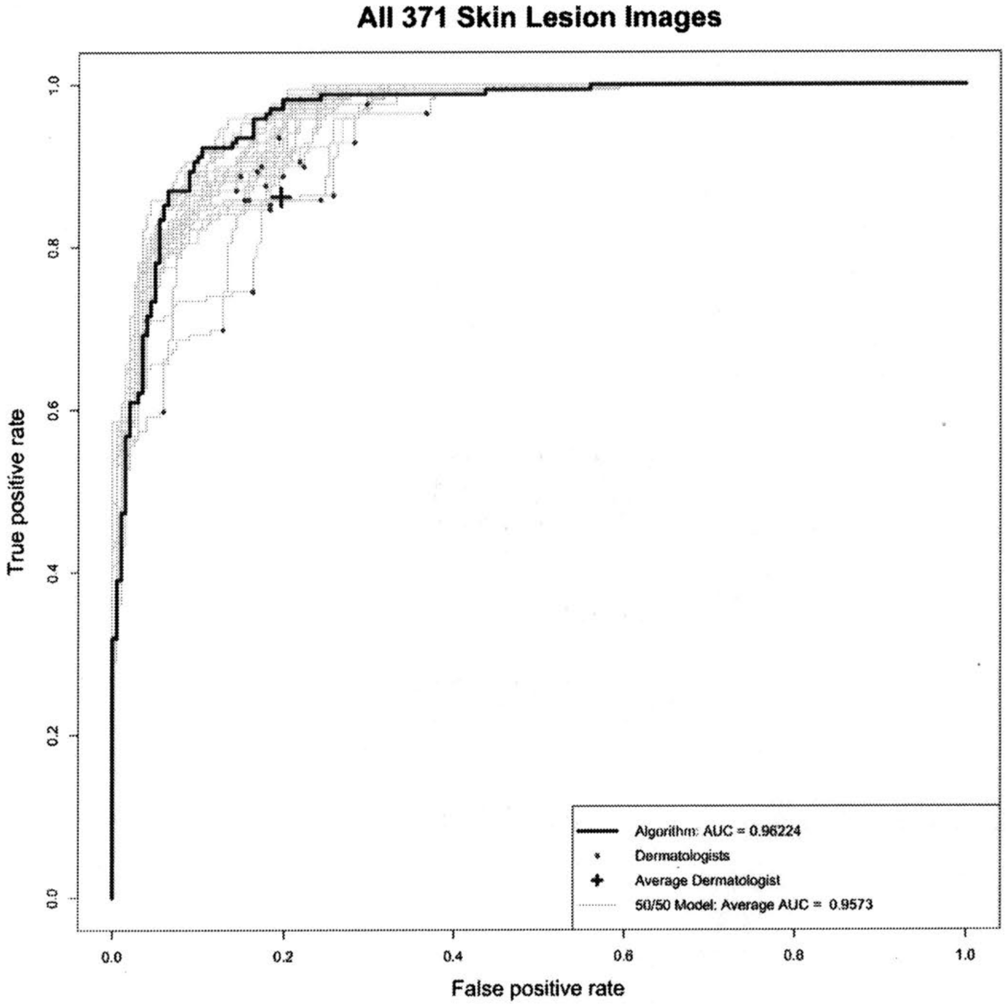


Figure 4. ROC curves of each of the 21 individual dermatologists when their opinions are averaged with the CNN (red lines) compared to the CNN alone (blue line).

In order to consider approaches where the CNN is given some degree of decision authority, Figure 5 shows all possible doctor/CNN weightings from 1-100%, rather than only the 50/50 model. As can be seen, for all 21 doctors there is some point at which the mixed human/AI model outperforms either alone, and for each doctor the optimal weighting scheme allows the CNN to have a weighting greater than 50%, thereby enabling it to sometimes overrule the doctor.

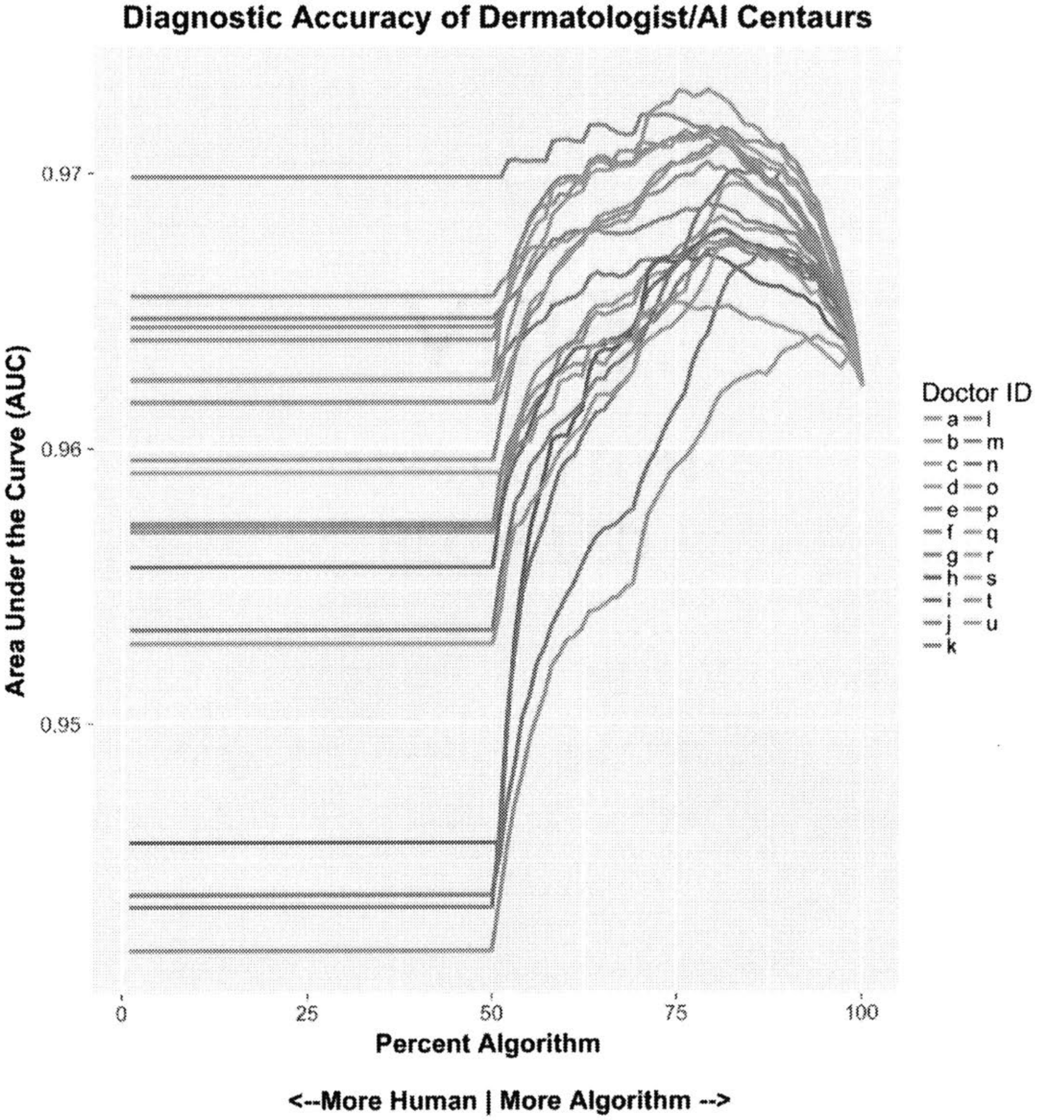


Figure 5. AUC achieved by combining each dermatologist's opinions with the CNN with weighting schemes from 99% dermatologist + 1% CNN, to 100% CNN. All models where the CNN is weighted 50% or below achieve the same AUC because the CNN is unable to overrule any dermatologist opinion, and for all 21 dermatologists the optimal weighting is >50% CNN.

Of course, doctors equipped with predictions from an artificial intelligence system could likely learn to allow the system to overrule their opinion under the right circumstances. As artificial intelligence decreases the cost of making accurate predictions, the relative importance of good human judgment—i.e., knowing how and when to act on those predictions—may be increasing (Agrawal, Gans, & Goldfarb, 2018). Indeed, the primary way in which artificial intelligence is currently utilized in healthcare is in Clinical Decision Support Systems (CDSS) such as Computer Aided Diagnosis (CAD) systems, which have rapidly entered mainstream medicine for tasks like mammography analysis (Shiraishi, Li, Appelbaum, & Doi, 2011). In these arrangements, a CAD system provides some initial input (e.g., a risk score, or identifying regions of interest) and human decision makers can decide how to incorporate this information into their diagnosis. However, a drawback of this approach is that it provides an opportunity for human decision makers to fall victim to a host of decision making biases. For instance, research has shown that people exhibit "algorithm aversion" and quickly lose trust in algorithms after seeing them err, even if they know that it acheives super-human performance and even if they see humans make the same mistake (Dietvorst, Simmons, & Massey, 2014). If doctors do not trust CAD systems, their potential to improve diagnostics dwindles. Despite their mainstream use in mammography, for instance, some studies suggest that CAD systems are of limited or even no value (Kohli & Jha, 2018; Lehman et al., 2015).

An alternative possibility for combining human and machine intelligence is to reverse the "standard partnership" and to have a computer incorporate human predictions into a final decision (Brynjolfsson & McAfee, 2017). Figure 6—which is constructed by randomly splitting the 371 cases into a training set of *n* images and a test set of 371-*n* images, fitting a simple linear model on the training set, calculating the AUC achieved by the model on the test set, and then repeating this process 1,000 times and averaging—

shows that it is possible to learn the optimal human/CNN weightings for each individual doctor within only about 100 cases or less. Over time, this approach would also make it possible to implement strategies for correcting human biases (even individual biases), which has been shown to improve predictive accuracy for making political judgments (Baron, Mellers, Tetlock, Stone, & Ungar, 2014; Mandel & Barnes, 2014).

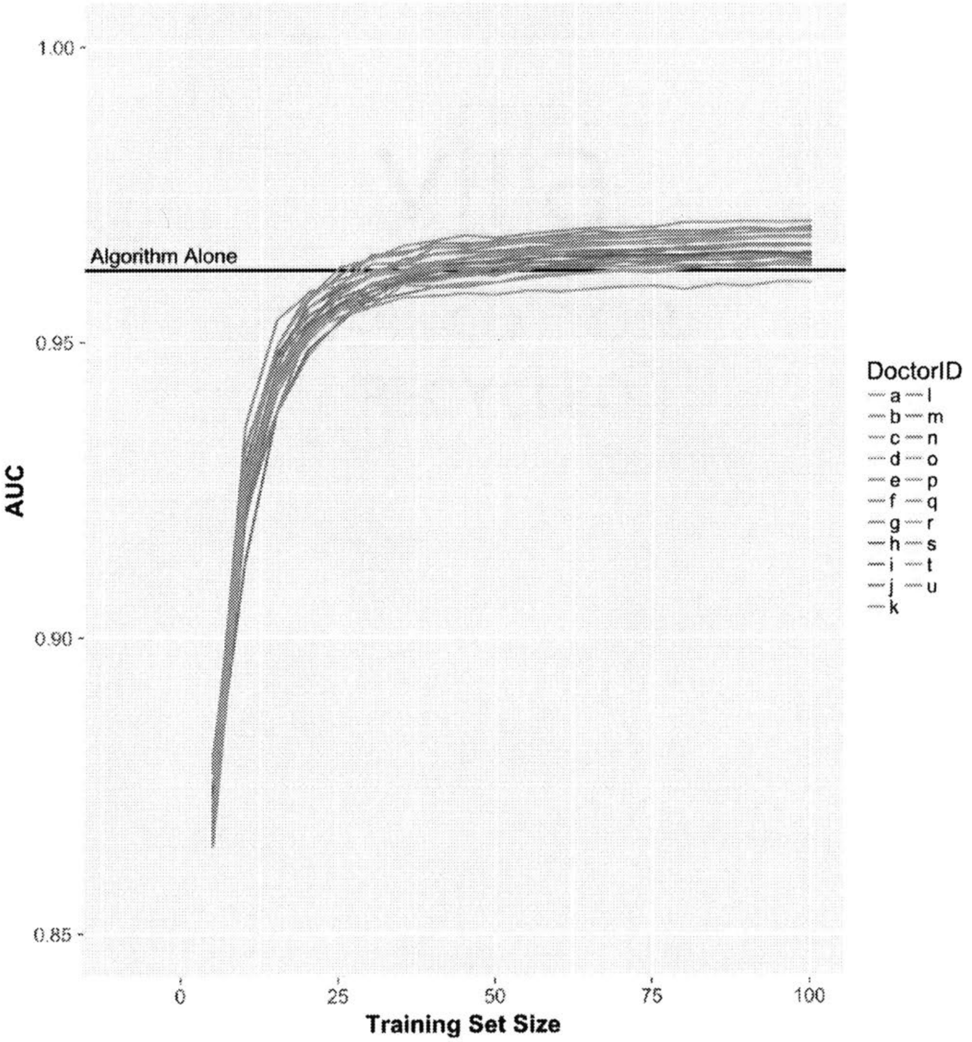


Figure 6. Learning curves for fitting linear models to combine each doctor with the CNN show that it only takes approximately 50-100 cases to learn a relative weighting where the dermatologist + CNN outperforms the CNN alone.

Reversing the standard partnership also allows for the same approach to be used for combining crowds with artificial intelligence. Figure 7 shows that as the number of dermatologists increases, the optimal weighting becomes less dependent on the CNN, and Figure 8 shows that these weightings can be learned in roughly 100 cases or less. In this case, because large groups of 20 dermatologists perform similarly to the CNN alone, the optimal weighting is close to 50/50.

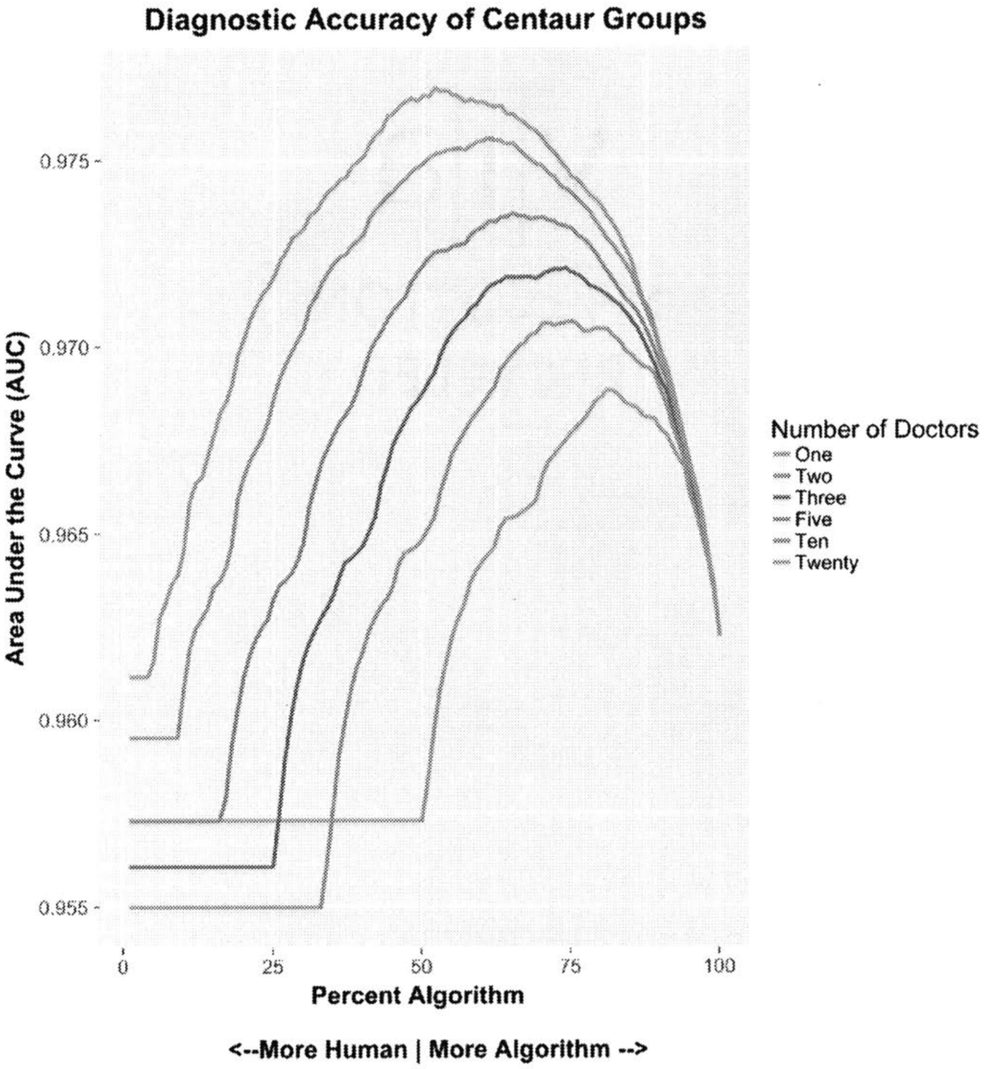


Figure 7. AUC achieved by combining crowds of dermatologists with the CNN with weighting schemes from (99% crowd + 1% CNN) to (100% CNN).

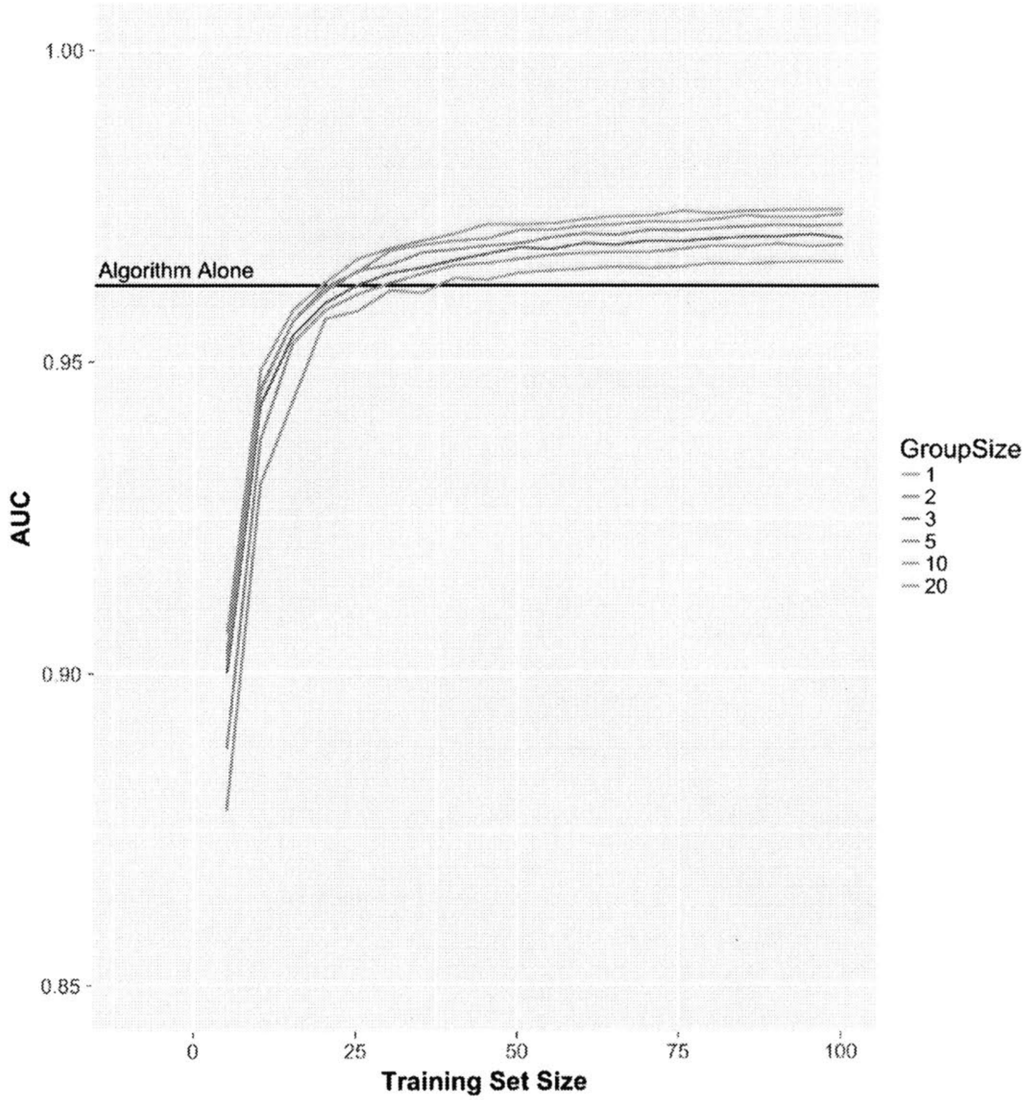


Figure 8. Learning curves for fitting linear models to combine crowds with the CNN show that it only takes approximately 50-100 cases to learn a relative weighting where the dermatologist + CNN outperforms the CNN alone.

Furthermore, even though linear aggregation methods are often a highly effective way to aggregate multiple diagnoses, it is possible that non-linear approaches—including the same machine learning methods used to create many artificial intelligence systems—could be used to further improve collective intelligence (e.g., Prelec, Seung, & Mccoy, 2017). One might think of this approach as "human ensemble learning." After all, expertise is in many cases multidimensional, and therefore over time one could potentially discover rich information about individual decision makers' "schools of thought," which could be used to dynamically solicit and combine opinions from the right people on a particular case, possibly even under cost constraints (Ertekin, Rudin, & Hirsh, 2014; Welinder, Branson, Belongie, & Perona, 2010). Thus, perhaps ironically, artificial intelligence can be used to more effectively identify and extract value out of people when it achieves super-human performance.

Finally, as Esteva et al. (2017) noted, the proliferation of smart phones could—if outfitted with deep neural networks—greatly extend the reach of dermatologists outside of the clinic and profoundly expand access to vital medical care. There is yet another reason to be optimistic: mobile devices and other information technologies also greatly expand the ease with which multiple opinions could be solicited on a given case, which can further improve diagnostic accuracy. Based on our findings here, artificial intelligence will not always replace human work, but rather, may transform how it is organized.

References

Agrawal, A., Gans, J., & Goldfarb, A. (2018). *Prediction Machines: The Simple Economics of Artificial Intelligence*. Boston, MA: Harvard Business Review Press.

Ariely, D., Au, W. T., Bender, R. H., Budescu, D. V., Dietz, C., Gu, H., Wallsten, T., Zauberman, G. (2000). The effects of averaging subjective probability estimates between and within judges. *Journal of Experimental Psychology: Applied*, *6*(2), 130–147.

Armstrong, J. S. (2001). Combining Forecasts. In J. S. Armstrong (Ed.), *Principles of Forecasting: A Handbook for Researchers and Practitioners*. Boston, MA: Springer.

Baron, J., Mellers, B. A., Tetlock, P. E., Stone, E., & Ungar, L. H. (2014). Two Reasons to Make Aggregated Probability Forecasts More Extreme. *Decision Analysis*, *11*(2).

Becker, A. S., Marcon, M., Ghafoor, S., Wurnig, M. C., Frauenfelder, T., & Boss, A. (2017). Deep Learning in Mammography. *Investigative Radiology*, *00*(00), 1.

Blattberg, R. C., & Hoch, S. J. (1990). Database Models and Managerial Intuition: 50% Model + 50% Manager. *Management Science*, *36*(8), 887–899.

Brynjolfsson, E., & McAfee, A. (2017). *Machine, Platform, Crowd: Harnessing Our Digital Future*. New York, NY: W.W. Norton & Company, Inc.,.

Clemen, R. T., & Winkler, R. L. (1986). Combining Economic Forecasts. *Source Journal of Business & Economic Statistics Journal of Business & Economic Statistics*, *4*(1), 39–46.

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics*, *44*(3), 837.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2014). Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err. *Journal of Experimental Psychology: General*, *144*(1), 1–12.

Ertekin, S., Rudin, C., & Hirsh, H. (2014). Approximating the Crowd. *Data Mining and Knowledge Discovery*, *28*(5–6), 1189–1221.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, *542*(7639), 115–118.

Galton, F. (1907). Vox populi. *Nature*, *75*, 450–451.

Hazlett, H. C., Gu, H., Munsell, B. C., Kim, S. H., Styner, M., Wolff, J. J., ... & Collins, D. L. (2017). Early brain development in infants at high risk for autism spectrum disorder. *Nature*, *542*(7641), 348.

Kohli, A., & Jha, S. (2018). Why CAD Failed in Mammography. *Journal of the American College of Radiology : JACR*, *15*(3 Pt B), 535–537.

Kurvers, R. H. J. M., Herzog, S. M., Hertwig, R., Krause, J., Carney, P. A., Bogart, A., Argenziano, G., Zalaudek, I., & Wolf, M. (2016). Boosting medical diagnostics by pooling independent judgments. *Proceedings of the National Academy of Sciences*, *113*(31), 8777–8782.

Lehman, C. D., Wellman, R. D., M Buist, D. S., Kerlikowske, K., A Tosteson, A. N., &

Miglioretti, D. L. (2015). Diagnostic Accuracy of Digital Screening Mammography With and Without Computer-Aided Detection. *JAMA Internal Medicine*, *175*(11), 1828–1837.

Long, E., Lin, H., Liu, Z., Wu, X., Wang, L., Jiang, J., … Liu, Y. (2017). An artificial intelligence platform for the multihospital collaborative management of congenital cataracts. *Nature Biomedical Engineering*, *1*, 0024.

Malone, T. W. (2004). *The Future of Work: How the New Order of Business Will Shape Your Organization, Your Management Style, and Your Life*. Boston, MA: Harvard Business School Press.

Mandel, D. R., & Barnes, A. (2014). Accuracy of forecasts in strategic intelligence. *Proceedings of the National Academy of Sciences*, *111*(30), 10984–10989.

Prelec, D., Seung, H. S., & Mccoy, J. (2017). A solution to the single-question crowd wisdom problem. *Nature*, *541*(7638), 532.

Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., … Ng, A. Y. (2017). CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. *ArXiv*. Retrieved from https://arxiv.org/pdf/1711.05225.pdf

Rogers, H. W. et al. (2015). Incidence estimate of nonmelanoma skin cancer (keratinocyte carcinomas) in the US population, 2012. *JAMA Dermatology*, *151*(10), 1081–1086.

Shiraishi, J., Li, Q., Appelbaum, D., & Doi, K. (2011). Computer-Aided Diagnosis and Artificial Intelligence in Clinical Imaging. *Seminars in Nuclear Medicine*, *41*(6), 449–462.

Somashekhar, S. P., Sepúlveda, M.-J., Norden, A. D., Rauthan, A., Arun, K., Patil, P., … Rohit, K. C. (2017). Early experience with IBM Watson for Oncology (WFO) cognitive computing system for lung and colorectal cancer treatment. *Journal of Clinical Oncology*, *35*(15), 8527–8527.

Surowiecki, J. (2005). *The Wisdom of Crowds*. New York, NY: Knopf Doubleday Publishing Group.

Welinder, P., Branson, S., Belongie, S., & Perona, P. (2010). The Multidimensional Wisdom of Crowds. *Advances in Neural Information Processing Systems*, 2424–2432.

Weng, S. F., Reps, J., Kai, J., Garibaldi, J. M., & Qureshi, N. (2017). Can Machine-learning improve cardiovascular risk prediction using routine clinical data? *PLoS ONE*, *12*(4), e0174944.

Wolf, M., Krause, J., Carney, P. A., Bogart, A., & Kurvers, R. H. J. M. (2015). Collective intelligence meets medical decision-making: the collective outperforms the best radiologist. *PloS One*, *10*(8), e0134269.

Acknowledgments

I thank Andre Esteva for sharing data, as well as his coauthors for their work in creating the dataset. I am also grateful to Thomas W. Malone for his input and advice throughout the project, and Erik Brynjolfsson for his helpful comments on an earlier draft. **Funding:** This research was supported in part by the Army Research Laboratory (grant number W911NF-09-2-0053). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed of implied, of the Army Research Laboratory of the U.S. Government.

This page intentionally left blank

ESSAY 2

Recruiting Hay to Find Needles: Recursive Incentives and Innovation in Social Networks

Erik P. Duhaime, Brittany Bond, Qi Yang, Patrick de Boer, & Thomas W. Malone

Abstract:

Finding innovative solutions to complex problems is often about finding people who have access to novel information and alternative viewpoints. Research has found that most people are connected to each other through just a few degrees of separation, but successful social search is often difficult because it depends on people using their weak ties to make connections to distant social networks. Recursive incentive schemes have shown promise for social search by motivating people to use their weak ties to find distant targets, such as specific people or even weather balloons placed at undisclosed locations. Here, we report on a case study of a similar recursive incentive scheme for finding innovative ideas. Specifically, we implemented a competition to reward individual(s) who helped refer Grand Prize winner(s) in MIT's Climate CoLab, an open innovation platform for addressing global climate change. Using data on over 78,000 CoLab members and over 36,000 people from over 100 countries who engaged with the referral contest, we find that people who are referred using this method are more likely than others both to submit proposals and to submit high quality proposals. Furthermore, we find suggestive evidence that among the contributors referred via the contest, those who had more than one degree of separation from a pre-existing CoLab member were more likely to submit high quality proposals. Thus, the results from this case study are consistent with the theory that people from distant networks are more likely to provide innovative solutions to complex problems. More broadly, the results suggest that rewarding indirect intermediaries in addition to final finders may promote effective social network recruitment.

1 INTRODUCTION

1.1 Open Innovation

Advances in communication technologies have made it easier than ever to harness the collective intelligence of large groups of people through crowdsourcing. For instance, crowdsourcing can be used to collect product opinions or valuable information about recent events and also to mobilize strangers toward collective action (Nichols & Kang, 2012; Mahmud et al., 2013; Savage, Monroy-Hernandez, & Höllerer, 2016).

Open innovation contests—a particularly impactful form of crowdsourcing—are designed to attract creative solutions to difficult problems from a large number of potential solvers (Murray, et al., 2012). By opening up a problem to thousands, millions, or eventually billions of people around the world, problem broadcasters can reap the benefits of vast distributed knowledge and also increase the chances that they will be able to find a "needle in a haystack" solution to a complex problem (Lakhani, Lifshitz-Assaf, & Tushman, 2013).

Importantly, research on open innovation contests has found that such contests work not only because problem broadcasters are provided with potential solutions from *more* people, but also because they receive many ideas from very *different* people. For instance, in analyzing a dataset on 12,000 scientists' activity in 166 contests on the open-innovation site Innocentive (www.Innocentive.com), Jeppesen and Lakhani (2010) found that the best ideas came both from people who are often socially marginalized, specifically women, and from people whose technical expertise was in a field different from the problem at hand (e.g., a chemist solving a biology problem). Relatedly, Duhaime, Olson, & Malone (2015) found that ideas submitted to MIT's Climate CoLab (www.climatecolab.org) —a website where anyone in the world can submit solutions to climate-related issues—were just as likely to come from women as from men, from residents of other countries as from the US, and from people without a graduate school education or experience with climate issues.

Since the advantages of open innovation contests over traditional means of innovating are driven by attracting not just more people, but more people with access to novel information and perspectives, a central question then becomes: what is the best way to recruit the best solvers?

1.2 Social Network Recruitment

Many organizations rely on social network recruitment for finding key individuals. For instance, employee referral programs are a common way for employers to tap into the potential value of current workers' social networks, and recruitment by word-of mouth is the most common method organizations use to attract job applicants (Marsden, 1994; Marsden & Gorman, 2001; Castilla, 2005). This literature has illuminated how "weak ties" (i.e., relationships to distant acquaintances rather than close associates) often enable access to novel information because they are more likely to bridge "structural holes" between social networks (Granovetter, 1974; Burt, 1992). Applying this insight to the domain of finding solvers for open innovation contests, we reasoned that activating weak ties would be especially important for recruiting the best solvers to an open innovation contest.

1.3. Motivating Weak Ties with Recursive Incentives

While weak ties may be the most important for successful social network recruitment, they can also be very difficult to motivate (Kim & Fernandez, 2017; Bond, Labuzova, & Fernandez, 2018). Furthermore, social network search often requires exploring several different paths of social links, and motivating weak ties may become increasingly difficult with increasing social distance (Bnaya et al., 2013). For instance, research on social search (i.e., finding specific distant individuals through successive network tie activation) has also highlighted the importance of weak ties and has shown that target individuals can be "found" several degrees of separation away, but such success is highly sensitive to the individual incentives of distant weak ties (Milgram, 1967; Dodds, Muhamad, & Watts, 2003).

One promising approach for motivating weak ties to boost social network recruitment was demonstrated by Pickard et al (2011), who won the 2009 DARPA Network Challenge in which teams competed to be the first to identify the locations of ten red weather balloons placed at ten different previously undisclosed locations around the United States (Pickard et al., 2011). The team implemented a "recursive incentive mechanism" where people were rewarded not only for identifying a balloon, but also for referring others who helped to identify a balloon.

The approach, depicted in Figure 1, had several desirable attributes that motivated people to spread the word about the contest. First, it reduced disincentives stemming from competition (i.e., the incentive to not tell others, lest they find the balloons instead). Second, it broadened the scope of people that one might refer to the contest. Whereas a traditional referral incentive would motivate people only to spread the word to people who they think might be directly interested, the recursive incentive scheme provided an incentive to share the contest with anyone who might be indirectly interested (i.e., interested in sharing the contest with others). In this way, the recursive incentive scheme incentivized people to spread word of the contest more broadly and also motivated intermediaries to continue spreading the word, enabling information about the contest to reach distant networks.

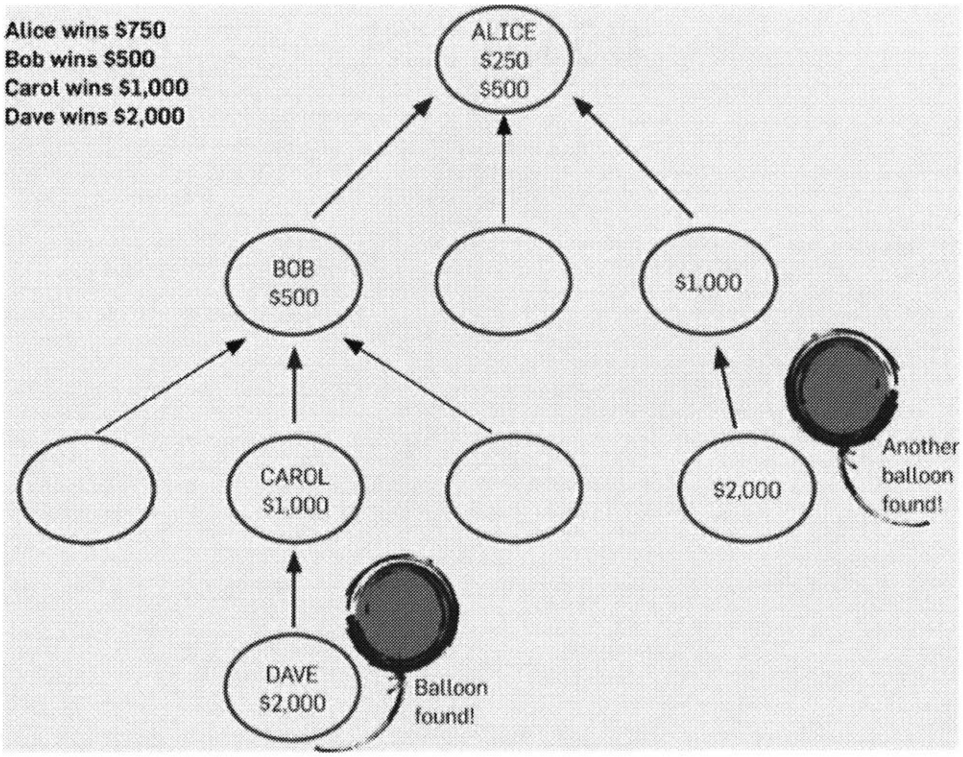


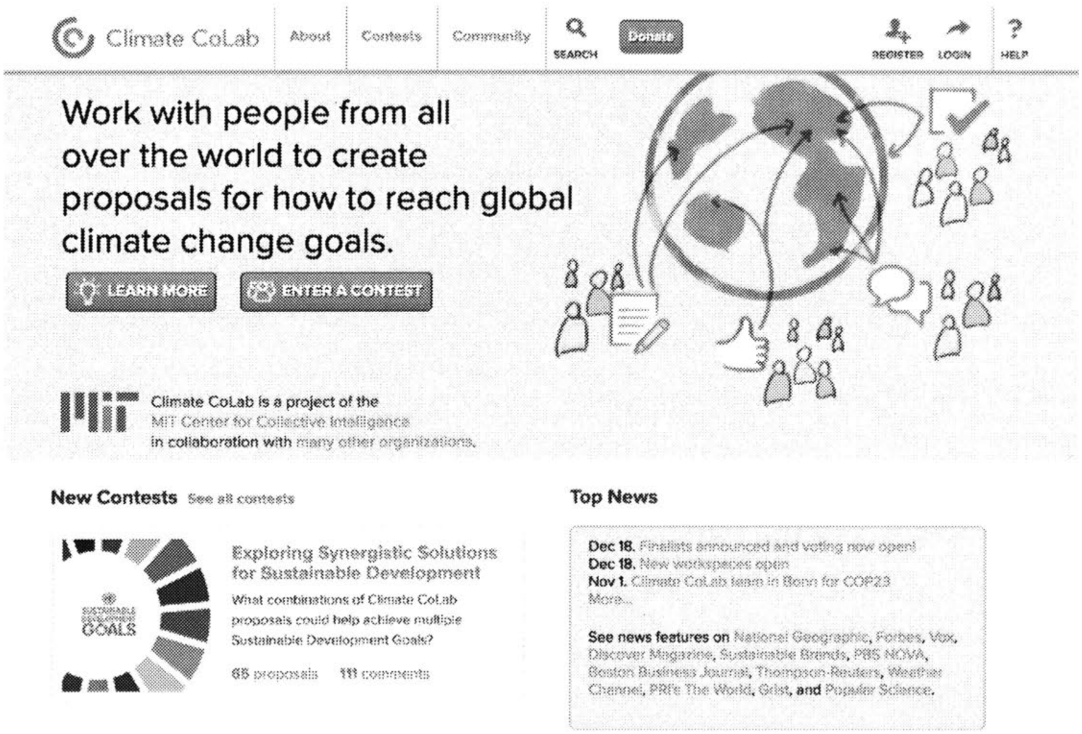
Figure 3: A figure used by the MIT Red Balloon Challenge team to explain the recursive incentive structure. Alice Joins the team and is given an invite link, like http://balloon.mit.edu/alice. Alice then emails her link to Bob, who uses it to join the team as well. Bob gets a unique link, like http://balloon.mit.edu/bob, and posts it on Facebook. His friend Carol sees it, signs up, then tweets about http://balloon.mit.edu/carol. Dave uses Carol's link to join, then spots one of the DARPA balloons. Dave is the first person to report the balloon's location to the MIT team, helping it win the Challenge. Once that happens, the team sends Dave $2,000 for finding the balloon. Carol gets $1,000 for inviting Dave, Bob gets $500 for inviting Carol, and Alice gets $250 for inviting Bob.

Here, rather than use a recursive incentive scheme to motivate people to find balloons, we used it to motivate people to find people with the best ideas for an open innovation contest. We reasoned that, like with the Red Balloon Challenge Team's approach, a recursive referral incentive would motivate people to spread word of the contest to weaker ties (i.e., not only people who might have an idea, but also to people who might know someone with an idea), thereby spreading word of the open innovation contest to more distant, heterogeneous social networks. In addition to bringing in more recruits, we hypothesized that many of these recruits would bring novel information and perspectives and as a result submit innovative ideas.

2 RESEARCH SETTING

2.1 MIT Climate CoLab

The MIT Climate CoLab (www.climatecolab.org) software platform allows individuals and teams of people from anywhere in the world to develop proposals for how to address global climate change (Malone & Klein, 2007; Introne, Laubacher, & Malone, 2011; Introne et al., 2011; Introne et al., 2013; Malone et al., 2017). As of July 2017, when data collection for this study ended, over 600,000 people from virtually every country in the world had visited the CoLab site, over 78,000 had registered as members, and over 1,250 had contributed to at least one proposal. (Today, there are over 100,000 members.)



Work with people from all over the world to create proposals for how to reach global climate change goals.

Climate CoLab is a project of the MIT Center for Collective Intelligence in collaboration with many other organizations.

New Contests

Exploring Synergistic Solutions for Sustainable Development

What combinations of Climate CoLab proposals could help achieve multiple Sustainable Development Goals?

65 proposals 111 comments

Top News

Dec 18. Finalists announced and voting now open!
Dec 18. New workspaces open
Nov 1. Climate CoLab team in Bonn for COP23
More...

See news features on National Geographic, Forbes, Vox, Discover Magazine, Sustainable Brands, PBS NOVA, Boston Business Journal, Thompson-Reuters, Weather Channel, PRI's The World, Grist, and Popular Science.

Figure 4: The MIT Climate CoLab homepage

Activity on the CoLab platform is driven by a series of annual contests (typically over a dozen) on topics such as how to reduce emissions in the transportation sector and how to change public attitudes about climate change. Some of the contests are run in conjunction with organizations such as the Union of Concerned Scientists (www.ucsusa.org) and the Carbon War Room (www.carbonwarroom.com). Each contest has advisors and judges, including experts from organizations like NASA, the World Bank, MIT, and Stanford, as well as one former US Secretary of State (Shultz), two former US Congress members (Inglis and Sharp), and two former heads of state (Robinson and Bruntland).

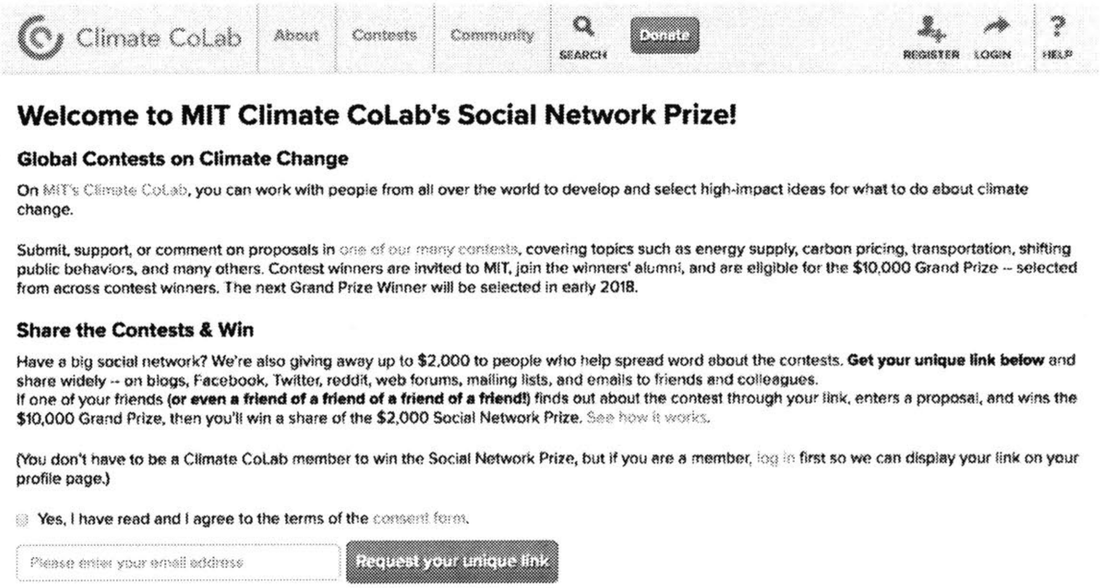
After proposals are submitted, the judges select the most promising entries to be semi-finalists, provide feedback to help improve the semifinalist proposals, and later select finalists. Then from the finalist proposals, the judges select the Judges' Choice Awards and the community votes for the Popular Choice Awards. The Climate CoLab offers a cash award—typically $10,000—to one Grand Prize Winner. All the Popular and Judges' Choice winners receive opportunities to present their ideas to top experts and potential implementers, usually at the Crowds & Climate conference, which was held at MIT annually from 2013-2016.

2.2 The Social Network Prize

Beginning with CoLab's annual contest of April 2014, we launched a referral contest with recursive incentives that we dubbed the "Social Network Prize." Each time someone clicked on a link for the competition they went to a landing page (Figure 3, below) explaining the contest. If a user clicked on the blue hyperlink to "see how it works," they were provided with an example based on the MIT Red Balloon Challenge Team's example presented in Figure 1. Specifically, it read:

> It might play out like this: Alice enters her e-mail address below and we give her a unique invite link. Alice then e-mails her link to Bob, who also enters his e-mail address below so that we can give him a unique invite link, which he then posts to Facebook. Bob's friend Carol sees the link, signs up to get her unique link, which she then posts to Twitter. Carol's follower Dave then sees the link, signs up for the CoLab and creates (or helps to create) the best-ranked proposal, winning him the Grand Prize of $10,000. Once that happens we send Carol $1,000 for inviting Dave, Bob gets $500 for inviting Carol, and Alice gets $250 for inviting Bob. If the chain of friends were even longer, then we would give out another $125, $67.50, and so on.

Upon reaching the landing page, if a person wanted to explore the CoLab site and become a member, they could easily do so. Even if they were not interested in becoming a member, they could still refer other people to the CoLab site and thus become eligible to win some of the Social Network Prize. To accomplish this, we needed their email address so that we would be able to give them their reward should they win it. Entering their email address generated a unique link associated with their ID number that they could send to people through Twitter, Facebook, e-mail, or by any other means.



Welcome to MIT Climate CoLab's Social Network Prize!

Global Contests on Climate Change

On MIT's Climate CoLab, you can work with people from all over the world to develop and select high-impact ideas for what to do about climate change.

Submit, support, or comment on proposals in one of our many contests, covering topics such as energy supply, carbon pricing, transportation, shifting public behaviors, and many others. Contest winners are invited to MIT, join the winners' alumni, and are eligible for the $10,000 Grand Prize -- selected from across contest winners. The next Grand Prize Winner will be selected in early 2018.

Share the Contests & Win

Have a big social network? We're also giving away up to $2,000 to people who help spread word about the contests. **Get your unique link below** and share widely -- on blogs, Facebook, Twitter, reddit, web forums, mailing lists, and emails to friends and colleagues.
If one of your friends **(or even a friend of a friend of a friend of a friend!)** finds out about the contest through your link, enters a proposal, and wins the $10,000 Grand Prize, then you'll win a share of the $2,000 Social Network Prize. See how it works.

(You don't have to be a Climate CoLab member to win the Social Network Prize, but if you are a member, log in first so we can display your link on your profile page.)

□ Yes, I have read and I agree to the terms of the consent form.

Figure 5: The landing page of the Social Network Prize contest. After agreeing to the consent form and entering their email address, people would be provided with a unique link that they could share with others. Once someone clicked on that link, they would arrive at this same landing page.

From an initial starter link for the competition as a whole, we created unique links for staff members to post on the CoLab website and to share in messages to all CoLab members, in blog posts, and in posts to various social media forums like Twitter, Facebook, and Reddit. Notably, staff members were not allowed to win the contest and therefore did not have a personal financial incentive to share with their networks. Thus, links were largely disseminated through the same channels that staff members typically used for other information about the CoLab. Staff members occasionally created and shared new links in order to disseminate them in different places, which would artificially inflate the size and complexity of the network. Therefore all links generated by a staff member were condensed into a unique staff referral link for analysis purposes.

Identifying unique chains of referral links was straightforward since everyone had to click on a unique link from someone else and then generate a unique link in order to share with others. However, we also needed to be able to identify whether a new Climate CoLab member was a Social Network Prize recruit (even if they did not themselves enter their email address to create their own unique link, which might happen if the new member was interested in the CoLab but not in recruiting new participants). Relatedly, when someone clicked on a Social Network Prize link, we needed to identify whether they were a new recruit or an existing member. We accomplished this by using browser cookies. Specifically, we installed a browser cookie when someone landed on the Social Network Prize landing page and/or were logged in to their CoLab account. That way, if someone was sent a Social Network Prize link and then became a member, we could associate the referral chain to the individual's membership. Only if the person was sent the referral link and later remembered the CoLab and Googled it on a different browser (i.e., instead of using the link) would they *not* be associated with their referral chain. Therefore determining whether a CoLab member who engaged with the Social Network Prize was an existing member or a new member recruited by the Social Network Prize entailed a simple comparison between the time the person first clicked on a Social Network Prize link and the time that they first created their membership.

While the Social Network Prize was re-launched in 2015 and 2016 and we include that data in our analysis, the majority of data for this case study comes from 2014 when staff made the most effort to spread the word of the contest.

3 DATA

3.1 The Referral Network

Over 36,000 people clicked on a Social Network Prize link and over 1,000 of those people entered their email address to generate their own unique link, which they could then share with their friends (a visualization of the spread of the Social Network Prize is depicted in Figure 4). Of the 36,000+ people who clicked on a Social Network Prize link, 351 people who were not already Climate CoLab members decided to join. The vast majority of these newly recruited members (309 of 351) were recruited by someone already directly affiliated with Climate CoLab—either by a current member or by a staff-generated link, but there were 42 new members who were indirectly recruited (i.e., by someone who was not a member when they clicked on the Social Network Prize link themselves). In other words, nonmembers who were also recruited by the Social Network Prize recruited these specific 42 new members.

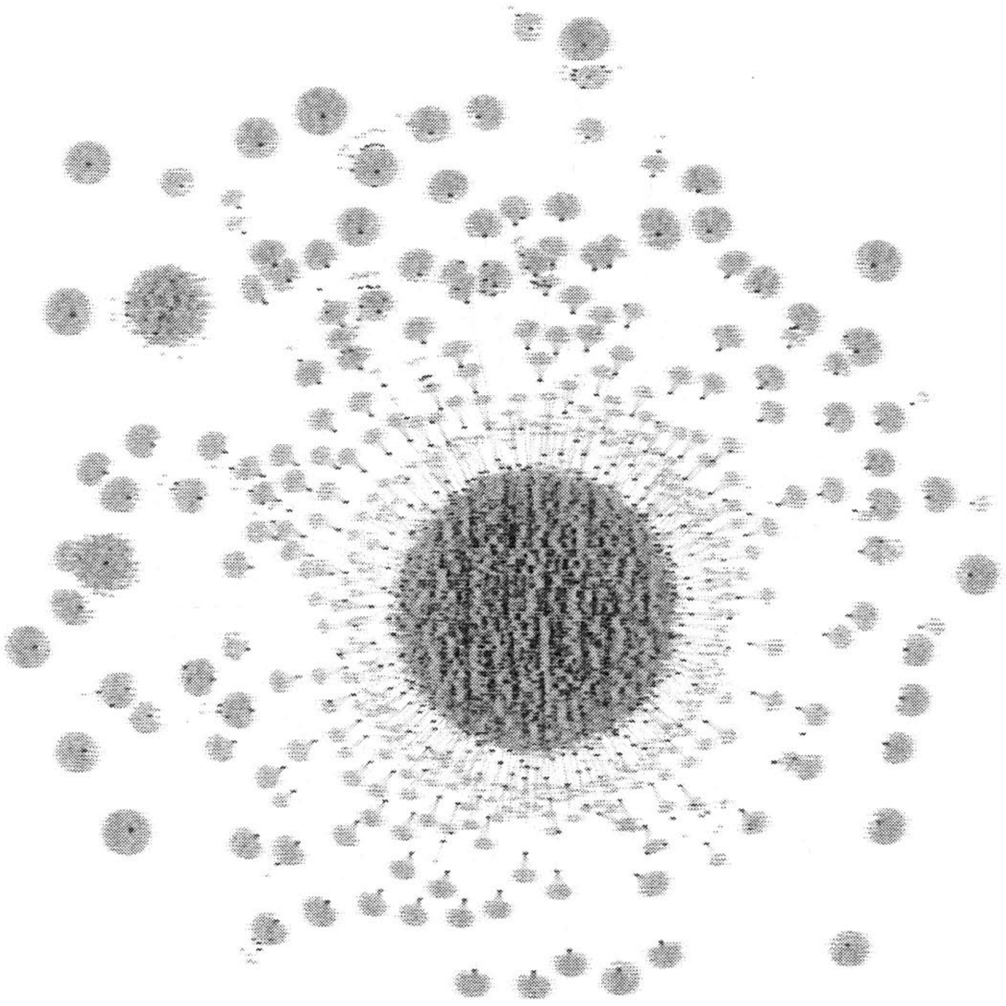


Figure 6: Visualization of all people who engaged with the Social Network Prize contest. Blue points indicate people who clicked on a Social Network Prize link. Red points indicate people who, upon clicking on a Social Network Prize link, entered their email address to create their own unique link that they could share with others. Lines connect a person to the person who shared the link with them, and the cluster in the center represents people who were one degree of separation from a CoLab member (i.e., people who were directly referred). The recursive incentive mechanism of the Social Network Prize provided an incentive for some people with whom the link was shared to share it with others as well, allowing it to spread further into other social networks.

3.2 Recruited Members Proposal Activity

The 351 newly recruited members made up a small portion (less than 1%) of the entire CoLab community of 78,390 members. However, as shown in Table 1, they contributed an outsized number of proposals: 57 of the 1284 proposal authors were newly recruited members from the Social Network Prize contests. Moreover, many of the proposals from these newly recruited members were high quality proposals that became finalists or even won their respective contests.

Table 1

	Users	**Proposal Authors**	**Finalists**	**Winners**
Social Network Prize New Member Recruits	351	57	16	9
→ Directly Recruited (i.e., referred by current member or staff-generated link)	309	52	13	7
→ Indirectly Recruited (i.e., referred by someone who was referred by someone)	42	5	3	2
Other CoLab Members	78039	1227	228	120
TOTAL CoLab Members	78390	1284	244	129

3.3 Participant Characteristics

For people who engaged with the Social Network Prize but did not become CoLab members, the only information we have is their IP address. From this information, we know that this group is highly geographically diverse: the 36,000+ people who clicked on a Social Network Prize Link were from over 100 different countries, the 1,000+ people who entered their email address were from over 60 different countries, and the 351 newly recruited members were from over 40 different countries.

We have more information on the people who were actually recruited through the Social Network Prize based on surveys sent to all CoLab members. The 2,607 unique survey responses from the entire CoLab community indicate that the community is male skewed (60%), has a median age of 30-39, is highly educated (55% attend or completed graduate school), and from all over the world (54% from outside of the United States). We have survey responses from 55 members who were recruited through the Social Network Prize, and these members were similar to the rest of the community in terms of gender (69% male), median age (30-39), and education (53% attend or completed graduate school). However, 80% of the newly recruited members live outside of the United States, which is significantly more than other survey respondents, $\chi^2(1) = 15.50$, $p < .001$.

While we do not have enough survey data on Social Network Prize recruits to make many meaningful comparisons between the characteristics of indirect vs. direct recruits, or finalists to non-finalists, it is notable that 8 finalists who were recruited by the Social Network Prize responded to the survey and all of them live outside of the United States.

4 RESULTS

4.1 Social Network Prize Recruits vs. Other CoLab Members

We first set out to compare the activity of CoLab members recruited via the Social Network Prize to that of other members. Whereas only 1.6% (1,227 of 78,039) of other members submitted proposals to CoLab contests, over 16% (57 of 351) of members recruited by the Social Network Prize did, which is a highly significant difference, $\chi^2(1) = 137353$, $p < .001$.

We next analyzed whether members recruited via the Social Network Prize were more likely to become finalists. While 4.6% of members recruited by the Social Network

Prize became finalists, only .003% of other members did, a highly significant difference, $\chi^2(1) = 130757$, $p < .001$.

We also analyzed whether members recruited via the Social Network Prize were more likely to become finalists conditional on having submitted a proposal. While 28.1% of authors recruited by the Social Network Prize (16 of 57) became finalists, only 18.6% of other authors (228 of 1227) became finalists. This difference is marginally statistically significant, $\chi^2(1) = 3.19$, $p = .074$.

4.2 Direct vs. Indirect Social Network Prize Recruits

We next set out to compare the activity of newly recruited CoLab members who were recruited directly by a member or a staff-generated link with those who were recruited indirectly (i.e., by someone who was also recruited via the Social Network Prize). While 16.8% (52 of 309) of directly-recruited members authored proposals, only 11.9% (5 of 42) of indirectly-recruited members did so. On account of the small sample sizes, we employed a Fisher's Exact Test instead of a chi-square and found that this difference was not statistically significant, $p = .509$.

Finally, we analyzed whether directly-recruited members were more likely to become finalists, conditional on having submitted a proposal, than indirectly-recruited members. While 25% of directly recruited authors became finalists (13 of 52), 60% of indirectly recruited proposal authors became finalists (3 of 5). Again employing a Fisher's Exact Test on account of the small sample size, we obtain a test statistic value of $p = .129$, providing modest evidence that indirect recruits author higher quality proposals than indirect recruits. Notably, two of the three finalist proposals from an indirect recruit won their respective contests (compared to 7 of 13 from direct recruits) and one of these was awarded the $10,000 Grand Prize in 2014, triggering financial awards for the Social Network Prize.

5 DISCUSSION

In our Social Network Prize case study, we find evidence that new members who were brought in by referrals were more likely to author proposals—and to author finalist proposals—than other CoLab members. We also find suggestive evidence that indirect recruits were more likely to author high quality proposals than direct recruits, which is consistent with the idea that innovative solutions to complex problems are likely to come from people in distant networks who bring in new information and perspectives. For instance, it is possible that by the time the Social Network Prize was launched, many pre-existing members had already told many of their close friends and colleagues with the best ideas for addressing climate change to join the CoLab, thereby exhausting the innovative potential of their local networks. However, the recursive incentive structure of the Social Network Prize opened up a new group of people who could be referred: people who might not have ideas themselves, but who might be likely to know other people with ideas. By incentivizing these weak ties, the Social Network Prize may have been able to mobilize not only more people than a traditional referral incentive, but also more people with new information and perspectives. This theoretical account is also supported by the survey data analysis, which indicates that users recruited from the Social Network Prize were significantly more likely to live outside the United States than other members.

Of course, our work has some limitations. Importantly, we only consider one particular case of leveraging a recursive incentive mechanism for promoting an open-innovation contest, and we do not know how well this result will generalize to other settings. We also have a relatively small sample size in terms of the number of new recruits, and especially the number of indirect recruits. Furthermore, we do not know enough about the characteristics of the recruits who authored finalist proposals to draw meaningful conclusions about them, although it is noteworthy that all eight for whom we have survey data live outside the United States. Therefore, while we know from other research that 1) recursive incentive mechanisms are a useful tool for accessing people with novel information and alternative viewpoints (Pickard et al., 2011) and 2) such people often submit innovative ideas to open innovation contests (Jeppesen & Lakhani, 2010; Duhaime, Olson, & Malone, 2015), we are unable to prove that this is why the proposals of Social Network Prize recruits were of high quality.

However, there are also reasons why our results may underestimate the potential effectiveness of recursive incentive schemes for social network recruitment. First, it can be difficult to quickly communicate how the recursive incentive scheme works because it is so much less common than traditional referral programs. But if recursive incentive schemes for social network recruitment become more common, this barrier to implementation will become less significant. Second, a desirable feature of the recursive incentive scheme is that it only costs an implementer more money than a traditional referral program *if it works*. This is precisely because of the recursive incentive structure. In other words, intermediaries are only compensated if they are critical for finding the end target. For these reasons, we believe that our findings are compelling enough that a wide range of organizations—especially those pursuing open innovation initiatives—might want to consider experimenting with referral programs that include recursive incentives.

REFERENCES

Bnaya, Z., Puzis, R., Stern, R., & Felner, A. (2013, September). Bandit algorithms for social network queries. In *International Conference on Social Computing* (pp. 148-153). IEEE.

Bond, B.M., Labuzova, T., & Fernandez, R.M. (2018). At the Expense of Quality. *Sociological Science*. 5: 380-411.

Burt, R. (1992). Structural Holes: The Social Structure of Competition. Cambridge, MA: Harvard University Press.

Castilla, E. J. (2005). Social Networks and Employee Performance in a Call Center. *American Journal of Sociology*. 110: 1243-83.

Dodds, P. S., Muhamad, R., & Watts, D. J. (2003). An Experimental Study of Search in Global Social Networks. *Science*, 301(5634), 827–829.

Duhaime, E. P., Olson, G. M., & Malone, T. W. (2015). Broad Participation in Collective Problem Solving Can Influence Participants and Lead to Better Solutions: Evidence from the MIT Climate CoLab. In *Proceedings of the 2015 Collective Intelligence Conference*. Santa Clara, CA.

Granovetter, M. (1974). The Strength of Weak Ties. *American Journal of Sociology* 78 (6): 1360-80.

Introne, J., Laubacher, R., & Malone, T. W. (2011). ROMA: A Framework to Enable Open Development Methodologies in Climate Change Assessment Modeling. *IEEE Software*, 28(6), 56–61.

Introne, J., Laubacher, R., Olson, G. M., & Malone, T. W. (2011). The Climate CoLab: Large scale model-based collaborative planning. In *Conference on Collaboration Technologies and Systems*. Philadelphia, PA.

Introne, J., Laubacher, R., Olson, G. M., & Malone, T. W. (2013). Solving Wicked Social Problems with Socio-computational Systems. *KI-Künstliche Intelligenz*, 27(1), 45–52.

Jeppesen, L. B., & Lakhani, K. R. (2010). Marginality and problem-solving effectiveness in broadcast search. *Organization science*, 21(5), 1016-1033.

Kim, M., & Fernandez, R. M. (2017). Strength matters: Tie strength as a causal driver of networks' information benefits. *Social Science Research*, 65, 268-281.

Lakhani, K. R., Lifshitz-Assaf, H., & Tushman, M. (2013). Open innovation and organizational boundaries: task decomposition, knowledge distribution and the locus of innovation. *Handbook of economic organization: Integrating economic and organizational theory*, 355-382.

Mahmud, J., Zhou, M. X., Megiddo, N., Nichols, J., & Drews, C. (2013). Recommending targeted strangers from whom to solicit information on social media. In *Proceedings of the 2013 international conference on Intelligent user interfaces*, 37-48. ACM.

Malone, T. W., & Klein, M. (2007). Harnessing collective intelligence to address global climate change. *Technology, Governance, Globalization*, 2(3), 15–26.

Malone, T. W., Nickerson, J. V., Laubacher, R. J., Hesse Fisher, L., de Boer, P., Han, Y., & Towne, B. W. (2017). Putting the Pieces Back Together Again: Contest Webs for Large-Scale Problem Solving. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (CSCW '17). Portland, OR.

Marsden, P. V. (1994). The hiring process: recruitment methods. *American Behavioral Scientist*, 37(7), 979-991.

Marsden, P. V., & Gorman, E. H. (2001). Social networks, job changes, and recruitment. In *Sourcebook of labor markets* (pp. 467-502). Springer US.

Milgram, S. (1967). The Small-World Problem. *Psychology Today*, 1(1), 61-67.

Murray, F., Stern, S., Campbell, G., & MacCormack, A. (2012). Grand Innovation Prizes: A theoretical, normative, and empirical evaluation. *Research Policy*, 41(10), 1779-1792.

Nichols, J. & Kang, J. (2012). Asking Questions of Targeted Strangers on Social Networks. *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, 999-1002

Pickard, G., Pan, W., Rahwan, I., Cebrian, M., Crane, R., Madan, A., & Pentland, A. (2011). "Time-Critical Social Mobilization". *Science*, 334(6055): 509-512.

Savage, S., Monroy-Hernandez, A., & Höllerer, T. (2016). Botivist: Calling volunteers to action using online bots. In *Proceedings of the 19th ACM Conference on*

Computer-Supported Cooperative Work & Social Computing (pp. 813-822). ACM.

ESSAY 3

Explaining the Decline of Tipping in the Gig Economy

Erik P. Duhaime & Zachary Woessner

Abstract:

Advances in information technology have enabled new ways of organizing work and led to a proliferation of what is known as the "gig economy." While much attention has been paid to how these new organizational designs have upended traditional employee-employer relationships, there has been little consideration of how these changes have impacted the social norms and expectations that govern the relationship between workers and consumers. Here, we consider the social norm of tipping and propose that gig-work is associated with a breakdown of tipping norms in part because of workers' increased autonomy in terms of deciding when and whether to work. We present four studies to support our hypothesis: a survey vignette experiment with workers on Amazon Mechanical Turk (Study 1), an analysis of New York City taxi data (Study 2), a field experiment with restaurant employee food delivery drivers (Study 3), and a field experiment with gig-worker food delivery drivers (Study 4). In Studies 1 and 2, we find that consumers are less likely to tip when workers have autonomy in deciding whether to complete a task. In Study 3, we find that restaurant delivery employees notice upfront tips (or lack thereof) and alter their service as a result. In contrast, in Study 4, we find that gig-workers who agree to complete a delivery for a fixed amount that includes an upfront tip (or lack thereof) are not responsive to tips. Together, these findings suggest that the gig economy has not only transformed employee-employer relationships, but has also altered the norms and expectations of consumers and workers.

INTRODUCTION

Advances in information technology have decreased the costs of coordination, leading to an overall shift towards proportionately more use of markets—rather than hierarchies—to coordinate economic activity (Malone, Yates, & Benjamin, 1987). In recent years, this shift has been most visible in what is often dubbed the "gig economy." In the gig economy, consumers can be seamlessly connected to on-demand independent contractor workers, rather than employees of a hierarchical organization. And unlike in traditional salaried jobs in hierarchical organizations that are typically paid by the year or by the hour, gig economy workers are paid by the task (Chen & Horton, 2016; Friedman, 2014).

A growing literature documents how the gig economy has upended the traditional employee-employer relationship and explores the implications for both workers and employers. The gig economy enables many workers to diversify their income streams through side ventures, such as performing freelancing work through websites like Upwork and driving for ridesharing companies like Uber and Lyft. Indeed, while many Americans do not report holding more than one job, more people are filing 1099 tax forms than in past years (Kuhn, 2016). For employers, the fact that these short-term workers do not need to be staffed year-round allows companies to avoid paying benefits like healthcare, and to better handle work fluctuations throughout the year (Houseman, 2001).

Less attention has been paid to how the growth of the gig economy has altered the social norms and expectations that govern the relationship between workers and consumers. Research in other domains has found that putting a price on a specific task—

a common feature of the gig economy—can lead to a breakdown of previously existing social norms. For instance, Ariely (2008) describes the success AARP experienced when asking lawyers for free services for needy retirees versus a discounted rate of $30—a fee which all the lawyers deemed too small. Zero dollars is rarely more attractive than $30, but "once market norms enter our considerations, the social norms depart" (Ariely, 2008). Similarly, Gneezy & Rustichini (2000) found that when day-care centers imposed a fine for parents who picked their children up late it led to *more* parents picking their children up late, presumably because the market mechanism of the fine crowded out previously existing social motivations. It stands to reason, then, that because consumers pay gig economy workers by the task, previously existing social motivations that governed the relationship between consumers and workers may deteriorate. Here, we consider one such social norm: tipping.

Tipping Norms in the Gig Economy

Tipping is a social phenomenon that generates some $42 billion dollars of income annually for workers in the American food industry alone (Azar, 2007, 2010). Some authors argue that tips serve as a pay-for-performance model to motivate worker performance (Lynn, Kwortnik, & Sturman, 2011; Shen, Ogawa, & Takahashi, 2014), while other common explanations for tipping are that people: 1) are altruistically motivated and tip primarily to "help servers" who make a low base wage, 2) tip to "reward service," driven by reciprocity norms and a desire to ensure that exchanges are equitable, and 3) have internalized tipping norms and tip out of a sense of duty or obligation (Harris, 1995; Lynn, 2015a, 2015b). Broadly, the strongest arguments for tipping combine extrinsic and intrinsic factors through both market transactions and social norms of reciprocity (Azar, 2003; Johnson, 2005). Meanwhile, when people fail to tip, research suggests that they are striving to avoid implied social status differences between themselves and the receiver, that they are signaling their displeasure with bad service, or that they simply want to save money (Lynn, 2015b).

With the advent of new mediums such as Uber, Lyft, and Grubhub for ridesharing and delivery services, consumer tipping behavior is changing commensurately. Several lines of evidence suggest that the gig economy is associated with a decline in tipping norms, and surveys indicate that young consumers who disproportionately contribute to the gig economy are less motivated by tipping norms than older consumers (e.g., Lynn, 2017). Customary tips for taxi drivers are over 20% of the base fare, whereas average tips to Uber drivers are reportedly only approximately 5% of the base fare (Wong, 2018).

Why is it that tipping taxi drivers is commonplace, but tipping Uber drivers is not? We propose the hypothesis that gig workers' increased autonomy over whether and when to work is a crucial factor. More specifically, because workers can choose when to work and are paid by the task—rather than by the year or the hour—market norms crowd out the internalized social tipping norms that compel some consumers to tip out of a sense of obligation or duty. Some consumers may also feel less motivated by reciprocity or equity norms, since in the gig economy a consumer can rest assured knowing that a worker would not have accepted the task in the first place if it were not worthwhile for them. These forces, in turn, decrease workers' expectations of tips and tip-based performance motivations. As a result, while many traditional service workers are motivated to "work for tips", we hypothesize that gig-economy workers are relatively motivated simply to complete tasks at the predetermined price they were offered.

We present four studies to support our hypothesis: a survey vignette experiment with workers on Amazon Mechanical Turk (Study 1), an analysis of New York City taxi data (Study 2), a field experiment with restaurant employee food delivery drivers (Study 3), and a field experiment with gig-worker food delivery drivers (Study 4).

STUDY 1

To test our theory that worker autonomy influences consumers' tipping motivations, we conducted a survey vignette experiment using workers on Amazon Mechanical Turk. We first collected pilot data with 174 participants that supported our hypothesis. We then collected a larger sample as described here to replicate our main result. In order to better understand participants' thinking, we also included an additional open-ended question that was not in the pilot study.

Method

Participants

We recruited 392 participants on Amazon Mechanical Turk for payment.

Materials and Procedure

Upon entering the survey, participants were asked to imagine that they needed groceries to make dinner later that night, but did not have any time to shop. They were then told about "a service, Shop4You, where you can pay for your groceries to be picked up and delivered to your house."

Manipulation

Participants were randomly assigned to one of two conditions: (1) a gig condition, or (2) an employee condition. After reading about the Shop4You service, participants in the gig condition read that "it will take approximately 1 hour to pick up and deliver your groceries, and the worker on Shop4You has agreed to do the job for $10." In contrast, participants in the employee condition read that "it will take approximately 1 hour to pick up and deliver your groceries, and Shop4You pays their employees $10 per hour."

Measures

Participants were then asked 1) whether they would tip the worker [yes/no], 2) how much they would tip the worker [$0-$10], and 3) why or why not they would tip the worker [free response].

Results

As shown in Figure 1, participants in the gig condition said that they would tip $2.53 (*SD* = 1.91) on average, significantly less than participants in the employee condition who said they would tip $3.06 on average (*SD* = 2.01), t(390) = 2.64, $p < .01$. Since the grocery delivery costs $10 in both conditions, this is an average tip of 25.3% in the gig condition vs. 30.6% in the employee condition. This difference was driven in part by the fact that more people in the gig condition say they would not tip at all. Whereas only 80% of participants said they would tip in the gig condition, 86.5% of participants said they would tip in the employee condition, a notable but not statistically significant difference, χ^2 (1) = 2.367, $p = .124$. However, participants who said they would tip in the employee condition still said they would tip more ($3.50 on average) than participants who said they would tip in the gig condition ($3.08 on average), t(325) = 2.18, $p = .030$.

Responses to the open-ended question support our hypothesis that worker autonomy is of critical importance. In the gig condition 38 of 192 participants said they would not tip at all, and 10 of those participants specifically mentioned the word "agree" in their explanation for their choice. For instance, participants who indicated that they would not tip said things like "they agreed to do it for $10 and nothing more," and "I do not believe this is a service that typically involves tipping. We have already agreed on a set fee amount." Four other participants did not specifically use the word "agree" but similarly attributed their decision not to tip to driver autonomy. For instance, one participant wrote, "from the information above, I would have ordered on-line and paid so picking up and delivery for $10 is appropriate *if the driver sets his own pricing*" [emphasis added]. Notably, these 14 people who specifically touched on worker autonomy represent 7% of participants in the gig condition, which is almost exactly the difference of non-tippers between the two conditions (86.5% vs. 80%).

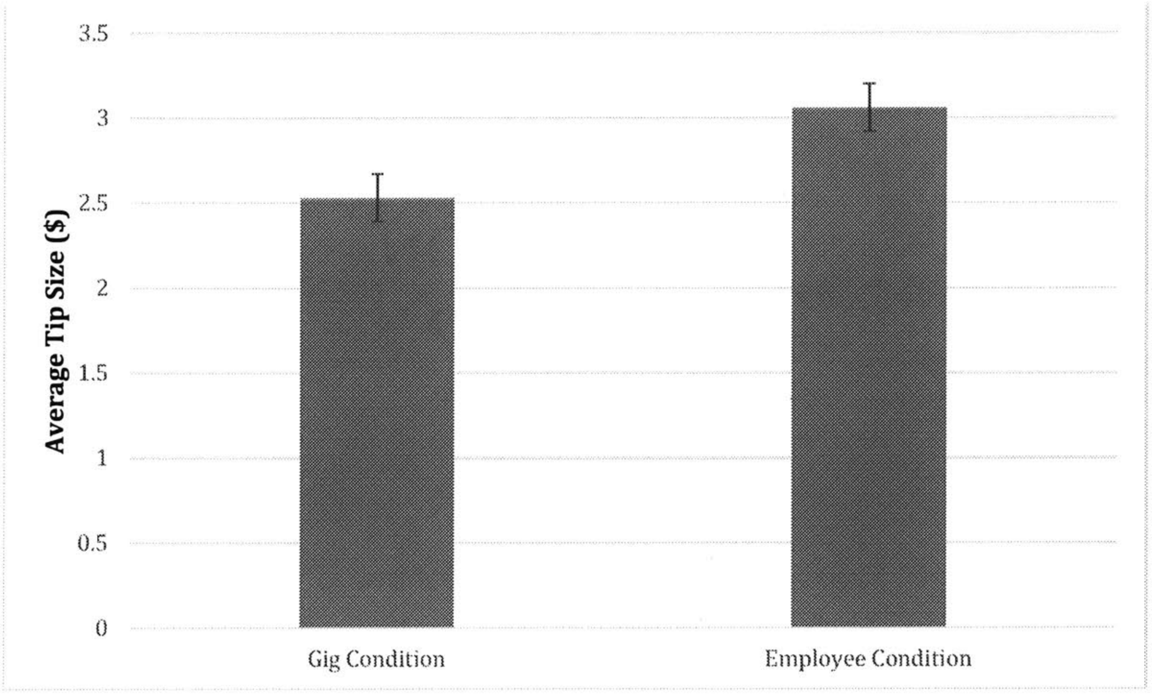


Figure 1: Differences in average tip size to hypothetical Shop4You workers when they are presented as gig workers vs. company employees. Error bars represent standard errors.

Discussion

The results support our hypothesis highlighting the role of worker autonomy in explaining the decline of tipping norms in the gig economy. While the survey only considered a hypothetical company, Shop4You, Study 1 shows that consumers' perceptions of worker autonomy can have a significant impact on consumers' tipping motivations.

STUDY 2

In an analysis of NYC taxi data published in *Bloomberg*, Chemi and Giorgi (2014) found that tip percentages fall sharply for fares ending in 5 or 0. To explain this "mysterious" effect, the authors surveyed a range of leading behavioral experts, including Richard Thaler from the University of Chicago, Dan Ariely from Duke, Andrew Lo from the Massachusetts Institute of Technology, and Cass Sunstein from Harvard, but none of the leading experts could offer a satisfying explanation. Despite not being able to explain the finding, Chemi and Giorgi suggested that a "trick for taxi drivers is to not let the fare hit that round number. The average tip at $60 is $8.82, but the average tip at $59 is $10.33. So in fact, going from $59 to $60 resulted in a loss of $1.50 in tip—more than the difference in fare."

We suspected that there might be something faulty with the conclusions of Chemi and Giorgi and, furthermore, that uncovering that error might reveal other rich insights about tipping behavior. More specifically, we suspected that the round number effect actually might be an artifact of the disproportionately large number of "Negotiated Flat Fare" rides that end in a 0 or 5, which are unmetered fares to locations outside of the city. According to the NYC Taxi and Limousine Commission[1], taxi drivers may choose whether to take such trips, and the fare must be mutually agreed upon by the driver and passenger before the trip may begin. Because drivers have autonomy over whether to accept an offer for a Negotiated Fare, we hypothesized that customers would feel less motivated to tip compared to when taking Metered Fares, where drivers are not able to reject a rider or set their own rate.

[1] http://www.nyc.gov/html/tlc/html/passenger/taxicab_rate.shtml

Data

Our data comes from the NYC Taxi and Limousine Commission dataset, which includes data on over a billion taxi rides in the last 10 years. To conduct our analysis, we select just one month of data—specifically, June 2013—comprising over 14 million rides. We remove all rides paid for in cash because cash tips are not included in the dataset, whereas tips paid for with a credit card are automatically included. To simplify the dataset, we also remove all fare types besides Metered Fares and Negotiated Fares, as well as rides with a fare amount of over $100, resulting in a dataset of 7,553,909 rides.

Results

Average tip size was $2.37 (*SD* = 1.95) on Metered Fare rides (*n* =7,525,174), compared to $6.46 (*SD* = 6.65) on Negotiated Fare rides (*n* = 28,735). However, since Metered Fare rides tend to be shorter and less expensive than Negotiated Fare rides out of the city, we regressed Tip Amount on Fare Type and controlled for Fare Amount in order to examine the effect of fare type on tipping behavior. Results, as seen in Figure 2, show that there is indeed a substantial difference in tipping behavior on Metered rides vs. Negotiated rides, and there does not appear to be a round number effect. As expected, we found a positive effect of Fare Amount, *b* = .176, *SE* = 0.000064, *t*(7553906) = 2757.7, *p* < .001, and a negative effect of Negotiated Fare Type, *b* = -2.32, *SE* = 0.0086, *t*(7553906) = 268.5, *p* < .001, indicating that tips to drivers for Negotiated Fare rides are significantly lower than those for Metered Fare rides.

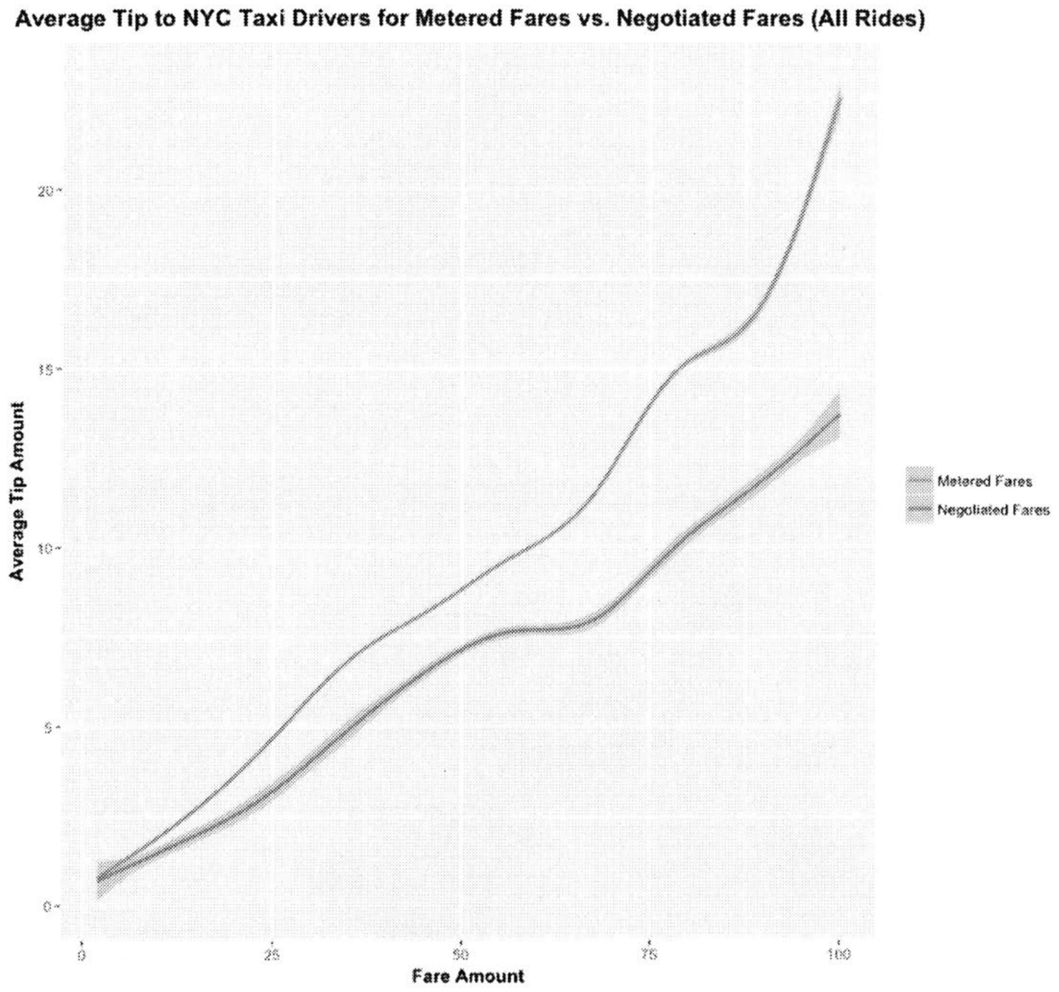


Figure 2: The impact of fare amount and fare type on average tip size (considering all passengers including non tippers) to NYC taxi drivers in the month of June 2013.

Next, we examined whether lower tips on Negotiated Fare rides were driven by fewer people tipping, or by people tipping less. We found that while 97.0% of people tipped on Metered Fare rides, only 72.4% of people tipped on Negotiated Fare rides. As shown in Figure 3, when non-tippers are excluded from the dataset the difference in tipping behavior between Metered Fares and Negotiated Fares shrinks, especially for fares below $50.

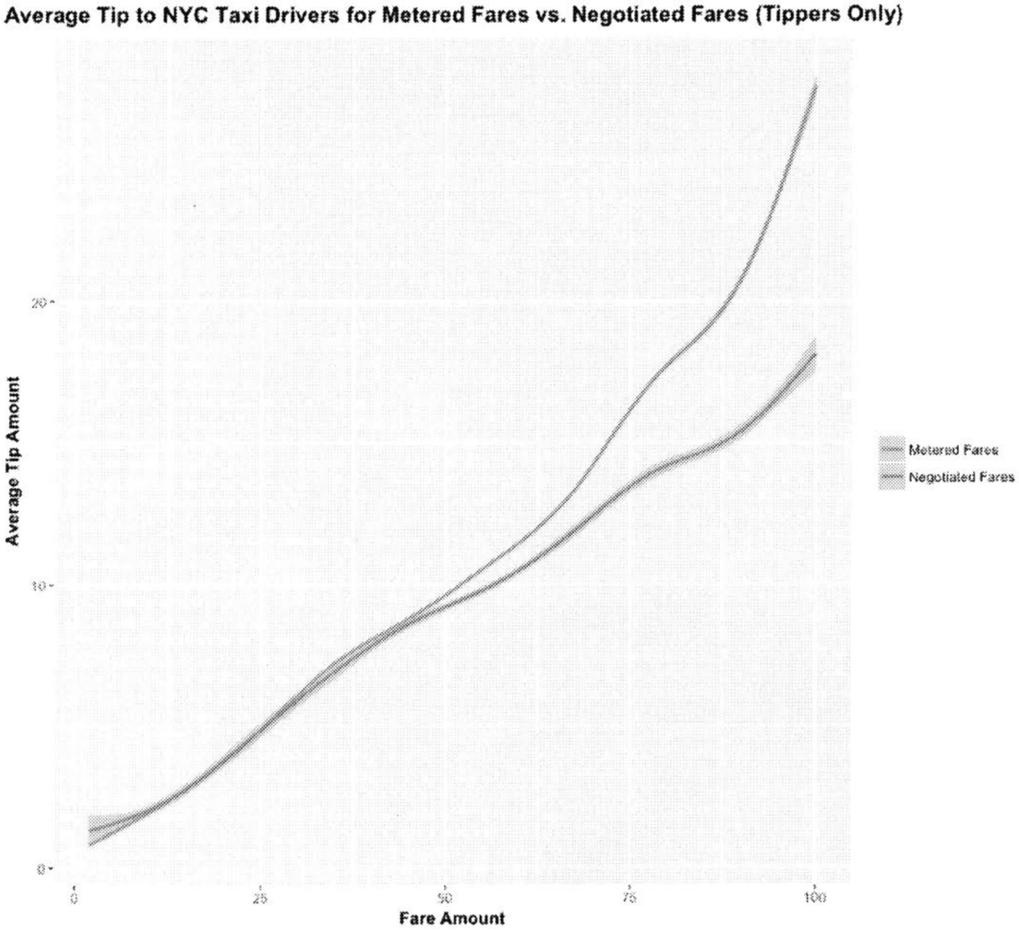


Figure 3: The impact of fare amount and fare type on average tip size (considering only those who tipped) to NYC taxi drivers in the month of June 2013.

Discussion

While taxi drivers are not typically considered gig workers, they effectively become so temporarily when they are hired for Negotiated Fares because they have autonomy over whether or not they complete these trips. This is an advantage of this study because a direct comparison of tips to taxi drivers vs. Uber drivers could be confounded by systematic differences in the types of people who choose to become one or the other (for instance, if passengers perceive that Uber drivers are wealthier than taxi drivers and less in need of tips) (Brewster, 2013, 2015). Here, the fact that people tip less, on average, when taking Negotiated Fare rides is consistent with our theory about the decline of tipping in the gig economy. Specifically, a crucial factor seems to be whether workers have autonomy in choosing whether or not to accept to complete a task at a predetermined price.

STUDY 3

Together, Studies 1 and 2 suggest that some consumers are less motivated to tip when workers have increased autonomy of whether to complete a task and for how much. How might such decreases in tipping norms, in turn, impact workers' expectations and behavior? To address this question, we ran two experiments in the domain of ordering food delivery. Whereas in the past consumers needed to call a restaurant directly to place an order, consumers can now place their orders with a credit card on a host of online platforms. As a result, tipping with a credit card at the time of ordering—rather than tipping in cash at the time of delivery—has become increasingly common. While some of these platforms simply take care of the ordering process, there are also many 3rd party delivery services where gig-workers pick up and deliver food so that the restaurants do not need to staff their own employees. In Studies 3 and 4, we explore the differences between these two models from the worker's perspective.

In Study 3, we ran an experiment with one such platform, Foodler, but did not order from any restaurants with 3rd party delivery drivers. The Foodler platform simply connected us to the restaurants and their employees and did not replace the delivery drivers, so presumably the restaurant employee drivers also complete many deliveries ordered the old fashioned way (i.e., over the phone). Therefore we hypothesized that these restaurant-employee delivery drivers would still expect tips, notice whether a tip has been provided up front, and alter their performance as a result. Specifically, we hypothesized that these delivery drivers would deliver food faster if they had not yet been tipped compared to when they were tipped before delivering the order. Furthermore, we hypothesized that if a driver noticed a relatively large upfront tip, they would be motivated by reciprocity norms to deliver food faster compared to when a small tip is provided upfront.

Method

Participants

Participants were 115 delivery drivers who were tasked with delivering food to an address in the greater Boston area over the course of 5 years (2012-2017).

Materials and Procedure

Sampling procedure was opportunistic: the experimenter and several of his friends collected a new data point each time they decided to order food over a 5-year period.

Manipulation

The experimenter manipulated whether the tip for the delivery was provided upfront or withheld until the time of delivery by alternating between the two. Partway through the experiment, the experimenter began alternating between no tip up front, a small tip up front (<15%) and a large tip up front (>25%) in order to increase variation of tip size.

Measures

The time the delivery took was the measure of interest, which was calculated based on the difference in time of the order confirmation email and the time of delivery. Other variables collected included the genre of the restaurant and the order subtotal.

Results

Orders were delivered in 38.77 minutes (*SD* = 13.21 minutes) when the tip was withheld until the time of delivery, which was significantly faster than the 45.67 (*SD* = 20.28 minutes) minutes on average when drivers were tipped up front, $t(79.8) = 2.22$, p =.036. We also found that when tips were provided upfront, drivers tended to deliver food faster when the tips were a larger percentage of the order subtotal (see Figure 4), despite seemingly not having any additional incentive to do so besides feelings of goodwill or reciprocity. Indeed, considering only those deliveries where tips were provided up front, regressing delivery time on tip percent yields a significant coefficient for tip percent, $b = -0.69$, $SE = .30$, $t(77) = 2.30$, $p = .024$, when also controlling for the genre of food delivered (Mexican, Pasta, Pizza, Sushi, Thai, or Wings).

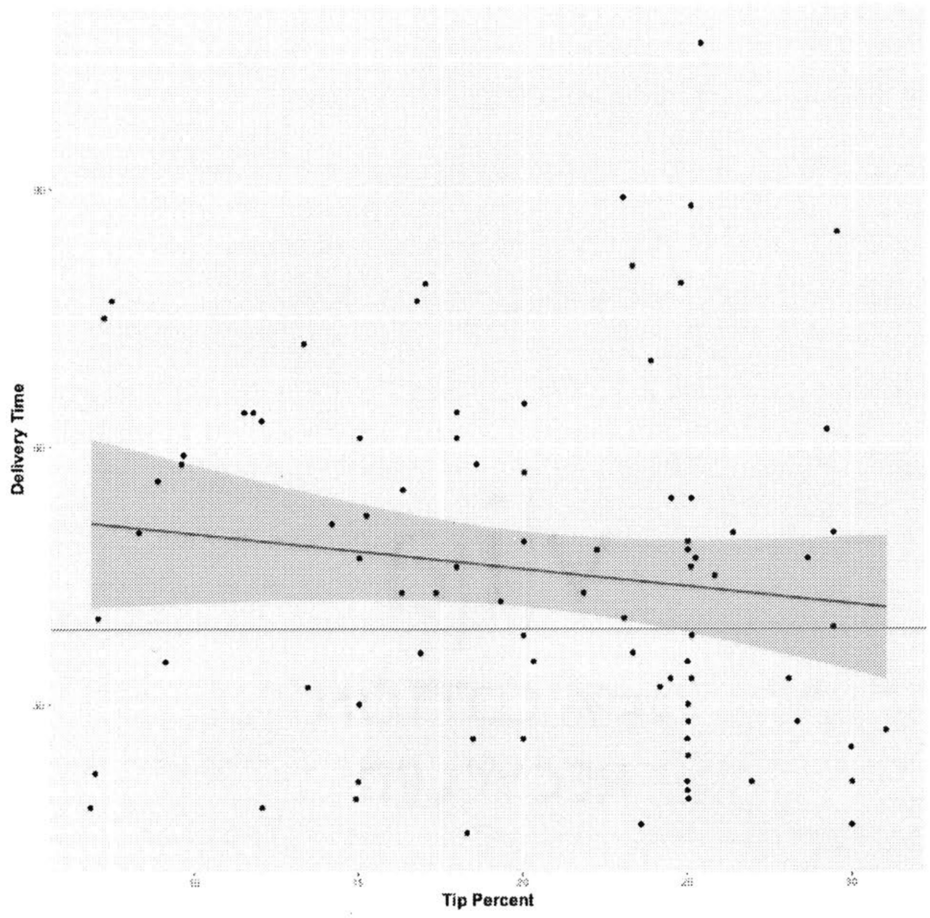


Figure 4: Orders were delivered faster, on average, when tips were withheld until the time of delivery (red line) compared to when they were provided upfront at the time of ordering (blue line). Furthermore, when tips were provided upfront, larger tip size led to somewhat faster deliveries.

Discussion

Study 3 shows that even when tasks are organized through on online platform, drivers still expect tips and alter their performance as a result when they are restaurant employees. Thus, Study 3 helps to isolate worker autonomy, rather than the use of 3rd party services for structuring work, as a driver of declining tipping norms. In other words, it suggests that tipping norms decline in the gig economy not because services are ordered online or over a smartphone app, but because the workers have control over whether they work at that time.

STUDY 4

While Study 3 shows that restaurant employee delivery drivers are motivated by tips, it does not show anything about how gig-workers in the same role may behave differently. To consider that question, we conducted a similar experiment in another city where Grubhub – another food ordering platform – was dominant. On Grubhub, drivers are usually independent contractors who do not work for the restaurant. As a result, the Grubhub drivers likely complete fewer deliveries ordered over the phone, and they likely complete fewer deliveries where a tip has not been provided upfront compared to the drivers in Study 3. Therefore, we hypothesized that Grubhub drivers have learned to expect fewer tips upon arrival at a customer's location and, as a result, are relatively less motivated by tips.

Furthermore, when Grubhub drivers accept the task of delivering an order, they see a composite price that incorporates the tip amount and specifies the pickup location (i.e., the restaurant address) and the drop-off location (i.e., the customer's location), and they then decide whether to accept or decline the delivery. Therefore, we hypothesized that compared to the drivers in Study 3, the Grubhub drivers would be less likely to even notice tips in the first place, and the size of upfront tips would have little impact on driver performance. Of course, the Grubhub drivers should still appreciate the additional income from larger tips. Rather, the difference is that the restaurant employees can only increase their wage by "working for tips," whereas the gig-workers can best increase their wage by completing as many high-value deliveries as quickly as possible regardless of the relative size of the tip compared to the base fee.

Method

Participants

154 food delivery orders were placed using Grubhub by 12 individuals, who were recruited by the experimenter in exchange for $6 per order. One of the 12 individuals collecting data (specifically, Z. Woessner), also interviewed drivers to learn more about their motivations and experiences. Interviews were conducted opportunistically with willing drivers, through asking friends to make connections to friends who had experience delivering food for Grubhub, and by cold calling and emailing the company.

Materials and Procedure

Orders were placed through online applications. Tips were randomly selected to be $0, $2, or $6 given on the app. Each driver received a total of $6; the difference between the randomly selected tips and $6 was made up in cash at the door.

Follow-up interviews were conducted with 5 Grubhub affiliates over the phone and online communication platforms, who we reached through cold calling and emailing the company. We used the following structured questions:

1. Can you see tips before accepting an order?
2. Do you expect tips when none are listed?
3. Does the size of the tip matter? If yes, please describe the different conditions.

4. When do customers tend to tip you?
5. Do you provide different service based on the tip?

Measures

Participants reported the time elapsed from placing the order until the food arrived at the door. Interviews with drivers provided qualitative data.

Results

The average delivery time across all conditions was 39.93 minutes (*SD* = 17.04 minutes). To determine if our experimental manipulation affected food delivery times, we first considered the average delivery time within each condition. In the $6 on the app condition, the average delivery time was 40.30 minutes; in the $2 on the app condition, it was 38.54 minutes; and in the $0 on the app condition it was 40.98 minutes. Next, in order to control for factors such as the subtotal of the order, the genre of food, and the distance to the restaurant, we regressed the total time of delivery on a dummy variable for the condition (i.e., $6 on the app, $2 on the app, or $0 at-the-door), the order subtotal, the distance (in miles) from the restaurant to the delivery address, and the genre of food. Of these, the only significant effect was the delivery distance ($b = 6.12$, $t=1.99$, $p=.0025$), indicating that each additional mile added, on average, 6 minutes to the delivery time. As shown in Figure 5, a simple regression of average time on tip percentage up front also found no effect, suggesting not only that the timing of tips did not matter, but also their size.

To further support this quantitative data, we conducted multiple, structured, qualitative interviews with Gruhub drivers and the company itself. We spoke with a Grubhub representative who confirmed that drivers can see the total amount of an order before accepting it. When speaking with the Grubhub drivers, we learned that this information is present, but the additional money they receive is presented as a bundle (delivery fee, mileage reimbursement, and tip). Drivers report using a rule of thumb to determine the tip (anything above $5 is tip), and that they do not expect any additional cash at the door for their service.

Drivers expressed value in being able to work at times that fit their schedule and location. The biggest complaint was not with tips given, but how Grubhub only allows workers to take jobs in blocks as opposed to individual deliveries. These blocks are roughly 2 hours long and require workers to sometimes take deliveries they otherwise would choose not to accept (for reasons of distance, value of the order compared to compensation, etc.).

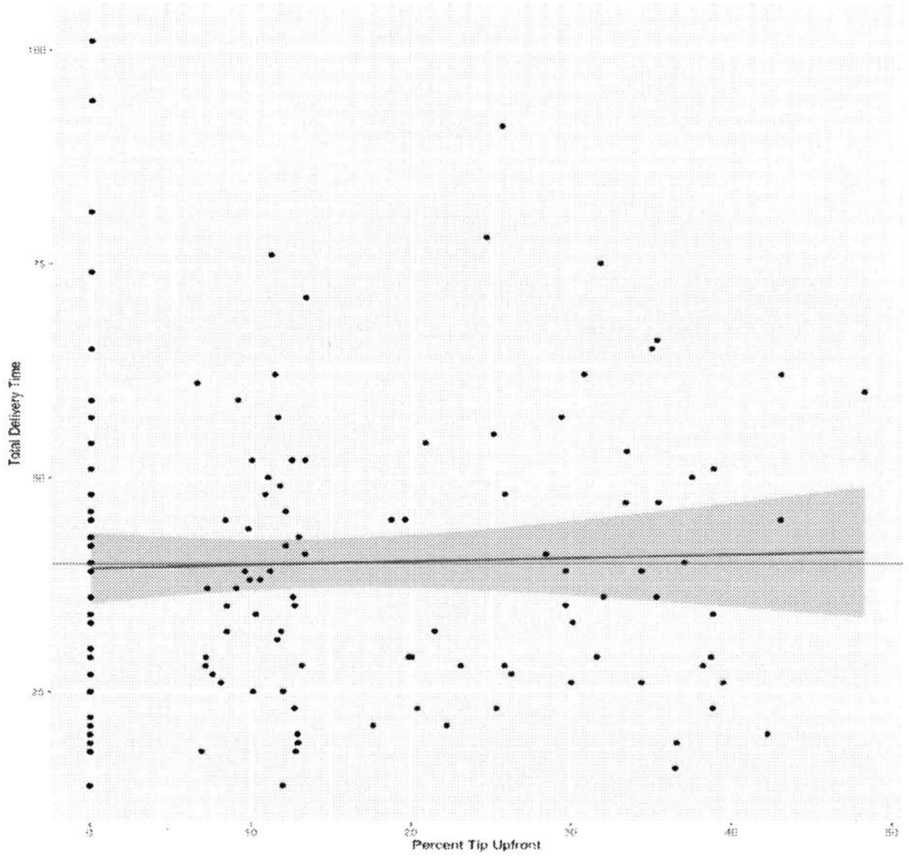


Figure 5: Orders were not delivered faster, on average, when tips were withheld until the time of delivery (red line) compared to when they were provided at the time of ordering. Furthermore, when tips were provided upfront (points) larger tip size did not lead to faster delivery time (blue line).

Discussion

The findings of Study 4 support our hypothesis that the size of the upfront tip will not impact delivery times when deliveries are made by gig-workers. While it is intriguing that larger upfront payments did not correlate with faster delivery times indirectly through increasing the total fee for those deliveries, it is perhaps not surprising given that Grubhub drivers are motivated to complete all rides as quickly as possible—regardless of the total fare—so that they can complete more total deliveries. In contrast, restaurant delivery drivers may find it advantageous to prioritize some deliveries over others in order to garner more and larger tips. Furthermore, the Grubhub drivers were reportedly surprised to receive any tips at the door, and often times almost left before the experimenters even had an opportunity to offer the cash tip. Thus, the Grubhub drivers have learned not to expect many tips at the door, and therefore they would have little incentive to prioritize some deliveries over others even if they were completing several orders at once.

GENERAL DISCUSSION

Study 1 demonstrates that people tip workers more if they perceive them to be employees, rather than gig workers who agree to complete a task at a set price. Study 2 extends this finding to a real world setting. Studies 3 and 4 provide evidence suggesting that tips have a significant impact on service when restaurant-employed drivers deliver food, but not when gig-workers do. Together, these studies suggest that worker autonomy over when and whether to work erodes customer motivations to tip and that gig workers are, in turn, relatively less motivated by tips than traditional service employees.

Of course, there are many limitations to theses studies. Broadly, one limitation is that the definition of gig work is inherently vague and shifting. For example, many workers fail to list their side jobs—such as driving for Uber—on survey data (Kuhn, 2016). That being said, the gig economy itself is described as "contingent work" by the Bureau of Labor Statistics (2017) and therefore, the presence of long-term gig workers might be discounted or overlooked by the fact that "contingent work" is often not considered to be a career. Another broad limitation of these studies is that we only study the impact of one social norm in just a few select contexts. Future work will examine the changes in tipping behavior in other settings, and examine other ways in which gig work models have transformed the relationship between consumers and workers.

More specifically, each of the four studies has limitations as well. In Study 1, perceptions of autonomy were not directly measured. While there are no differences between the two key conditions except for the framing of the workers as having autonomy over whether to complete the task, the results may actually be driven by some other association, such as a belief that gig-workers have access to other income streams and are less in need of tips. Similarly, in Study 2, while all the workers are NYC taxi drivers it is possible that some drivers are more likely to accept Negotiated Fares, and these drivers may have other differences that make them less likely to receive tips. Studies 3 and 4 explore workers' responses to tips as employees and as gig workers. However, while the results are consistent with the idea that tipping norms have changed as a result of the switch from an employee-driver model to a gig-driver model, there many also be other factors explaining the differences observed, such as the fact that the studies were conducted in two different cities.

Despite these limitations, these four studies together paint a consistent picture of how tipping norms are evolving in the gig economy. Importantly, once tipping norms deteriorate, it may be very difficult to re-establish them (e.g., see Gneezy and Rustichini, 2000). When looking at a company like Uber, a large employer of contract workers, consumers have received the same message for years: "There's no need to tip" (Rosenbloom, 2016). Although it was formerly said that the tip is included, that is no longer the case. For instance, an article titled "Uber's New Tipping Policy Is a Mistake" (Mohammed, 2016) outlines the changes Uber made to its tipping policy, essentially punctuating a seamless rideshare experience with a murky mix of social norms and market transactions. Signs were often hung explaining that "tips are not included on Uber's platforms," but that "riders are free to offer tips and drivers are free to accept them" (Mohammed, 2016).

More broadly, an implication of these studies is that while the marketization of tasks may make things more efficient, it may also have the unforeseen effect of crowding out preexisting social norms and expectations. Whether this is for better or worse depends on the context and one's perspective. Certainly some social norms may have more costs than benefits and, indeed, tipping norms have many negative consequences such as enabling a form of racial wage discrimination (Lynn et al. 2008). But it may also be that by eroding previously existing social norms, the marketization and gig-ification of tasks increasingly enables both managers and customers to view workers as another means of production, rather than as fellow human beings.

REFERENCES

Ariely, D. (2008). *Predictably Irrational*. New York: Harper Collins.

Azar, O. H. (2003). The implications of tipping for economics and management. *International Journal of Social Economics*, *30*(10), 1084–1094.

Azar, O. H. (2007). Do people tip strategically, to improve future service? Theory and evidence. *Canadian Journal of Economics*, *40*(2), 515–527.

Azar, O. H. (2010). Do people tip because of psychological or strategic motivations? An empirical analysis of restaurant tipping. *Applied Economics*, *42*(23), 3039–3044.

Berkhout, E., Heyma, A., & Prins, J. (2013). flexibility @ work 2013: yearly report on flexible labor and employment.

Brewster, Z. W. (2013). The effects of restaurant servers' perceptions of customers' tipping behaviors on service discrimination. *International Journal of Hospitality Management*, *32*, 228–236.

Brewster, Z. W. (2015). Perceptions of intergroup tipping differences, discriminatory service, and tip earnings among restaurant servers. *International Journal of Hospitality Management*, *46*, 15-25.

Chemi, E., & Giorgi, A. (2014). The Three Unexplained Mysteries of Taxi Tipping Behavior. *Bloomberg*. Retrieved from http://www.bloomberg.com/bw/articles/2014-08-07/tipping-taxi-drivers-data-analysis-cant-explain-these-puzzles

Chen, D. L., & Horton, J. J. (2016). Are online labor markets spot markets for tasks? A field experiment on the behavioral response to wage cuts. *Information Systems Research*, *27*(2), 403.

Friedman, G. (2014). Workers without employers: Shadow corporations and the rise of the gig economy. *Review of Keynesian Economics*, 2(2), 171-188.

Gneezy, U., & Rustichini, A. (2000). A fine is a price. *Journal of Legal Studies*, *29*(1).

Gramm, C. L., & Schnell, J. F. (2001). The use of flexible staffing arrangements in core production jobs. *Industrial and Labor Relations Review*, *54*(2), 245–258.

Harris, M. B. (1995). Waiters, Customers, and Service: Some Tips About Tipping. *Journal of Applied Social Psychology*, *25*(8), 725.

Houseman, S. N. (2001). Why Employers Use Flexible Staffing Arrangements: Evidence from an Establishment Survey. *ILR Review*, *55*(1), 149–170.

Johnson, C. (2005). Employee motivation: A comparison of tipped and non-tipped hourly restaurant employees.

Kuhn, K. M. (2016). The Rise of the "Gig Economy" and Implications for Understanding Work and Workers. *Industrial and Organizational Psychology*, *9*(1), 157.

Department of Labor (2017). Contingent and Alternative Employment Arrangements. *Bureau of Labor Statistics.* https://www.bls.gov/news.release/conemp.nr0.htm

Lynn, M. (2015a). Explanations of service gratuities and tipping: Evidence from individual differences in tipping motivations and tendencies. *Journal of Behavioral and Experimental Economics*, *55*, 65–71.

Lynn, M. (2015b). Service gratuities and tipping: A motivational framework. *Journal of Economic Psychology*, *46*, 74–88.

Lynn, M. (2017). Should U.S. restaurants abandon tipping? A review of the issues and evidence. *Psychosociological Issues in Human Resource Management,* 5(1), 120-159.

Lynn, M., Kwortnik Jr, R. J., & Sturman, M. C. (2011). Voluntary tipping and the selective attraction and retention of service workers in the USA: An application of the ASA model. *The International Journal of Human Resource Management*, 22(9), 1887-1901.

Lynn, M., Sturman, M., Ganley, C., Adams, E., Douglas, M., & McNeil, J. (2008). Consumer racial discrimination in tipping: A replication and extension. *Journal of Applied Social Psychology*, *38*(4), 1045-1060.

Malone, T. W., Yates, J., & Benjamin, R. I. (1987). Electronic Markets and Electronic Hierarchies. *Communications of the ACM*, *30*(6), 484–497.

Mohammed, R. (2016). Uber's New Tipping Policy Is a Mistake. *Harvard Business Review*.

Rosenbloom, S. (2016). To Tip or Not to Tip Your Uber Driver. *The New York Times*. https://www.nytimes.com/2016/05/22/travel/uber-taxi-tipping.html

Shen, J., Ogawa, K., & Takahashi, H. (2014). Examining the tradeoff between fixed pay and performance-related pay: A choice experiment approach. *Review of Economic Analysis*, *6*(2), 119–131.

Wong, K. (2018). Should You Tip Your Uber Driver? If So, How Much? *The NewYork Times*. Retrieved from https://www.nytimes.com/2018/10/02/travel/should-you-tip-your-uber-driver-if-so-how-much.html