

Human–Computer Interaction Series

Vassillis-Javed Khan
Konstantinos Papangelis
Ioanna Lykourantzou
Panos Markopoulos *Editors*

Macrotask Crowdsourcing

Engaging the Crowds to Address
Complex Problems

 Springer

Human–Computer Interaction Series

Editors-in-Chief

Desney Tan
Microsoft Research, Redmond, WA, USA

Jean Vanderdonckt
Louvain School of Management, Université catholique de Louvain,
Louvain-La-Neuve, Belgium

The Human–Computer Interaction Series, launched in 2004, publishes books that advance the science and technology of developing systems which are effective and satisfying for people in a wide variety of contexts. Titles focus on theoretical perspectives (such as formal approaches drawn from a variety of behavioural sciences), practical approaches (such as techniques for effectively integrating user needs in system development), and social issues (such as the determinants of utility, usability and acceptability).

HCI is a multidisciplinary field and focuses on the human aspects in the development of computer technology. As technology becomes increasingly more pervasive the need to take a human-centred approach in the design and development of computer-based systems becomes ever more important.

Titles published within the Human–Computer Interaction Series are included in Thomson Reuters’ Book Citation Index, The DBLP Computer Science Bibliography and The HCI Bibliography.

More information about this series at <http://www.springer.com/series/6033>

Vassillis-Javed Khan · Konstantinos Papangelis ·
Ioanna Lykourantzou · Panos Markopoulos
Editors

Macrotask Crowdsourcing

Engaging the Crowds to Address Complex
Problems

 Springer

Editors

Vassillis-Javed Khan
Eindhoven University of Technology
Eindhoven, The Netherlands

Konstantinos Papangelis
Xi'an Jiaotong-Liverpool University
Suzhou, China

Ioanna Lykourentzou
Department of Information
and Computing Sciences
Utrecht University
Utrecht, The Netherlands

Panos Markopoulos
Department of Industrial Design
Eindhoven University of Technology
Eindhoven, The Netherlands

ISSN 1571-5035

ISSN 2524-4477 (electronic)

Human-Computer Interaction Series

ISBN 978-3-030-12333-8

ISBN 978-3-030-12334-5 (eBook)

<https://doi.org/10.1007/978-3-030-12334-5>

© Springer Nature Switzerland AG 2019

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface: Macrotask Crowdsourcing for Advancing the Crowd's Potential

Amazon's launch of Mechanical Turk (MTurk.com) in 2005 kickstarted a new socio-technical phenomenon and a new labor model—that of crowdsourcing. Nowadays, MTurk is just one of several crowdsourcing platforms (à Campo et al. 2018). Such platforms bring two groups of people together: people who request a certain task but lack the skill or the time or the human capital to complete it, aka the *task requesters*; and people who wish to work on such tasks typically for a monetary reward, aka the *crowdworkers*.

The introduction of Application Programming Interfaces (APIs) in crowdsourcing platforms made the process of requesting and completing tasks much easier. These APIs have enabled the emergence of new scientific fields which integrate human effort with computing systems. *Human computation* is one of these fields, which channels human intelligence through the use of computing systems to solve tasks that no known efficient algorithm can yet solve (Von Ahn 2008). *Collective Intelligence* is another neighboring field, which couples human and machine intelligence to solve complex problems which neither humans nor machines can solve on their own (Malone et al. 2010).

The adoption of such platforms from large numbers of both requesters and workers, and the introduction of APIs established crowdsourcing as a fertile ground for researchers. However, due to the fact that widely adopted platforms like MTurk only supported short and easy tasks, known as microtasks (from the Greek word *μικρός*, which means small) research studies so far and most industrial applications have primarily focused on microtask crowdsourcing. To a certain extent, this focus is rightful; microtask crowdsourcing has produced some very impressive results. Examples include labeling images for improving image search and web accessibility (Von Ahn & Dabbish 2004); editing documents for shortening and proof-reading (Bernstein et al. 2010); captioning audio in real time for accessibility (Lasecki et al. 2012); getting feedback on articles (Kittur et al. 2008) and designs (Luther et al. 2014). Consequently, most industrial practitioners and researchers today, when thinking of crowdsourcing they automatically think of a large list of small, similar, homogeneous and relatively straightforward to complete tasks—i.e., microtasks.

But not all types of work can be accomplished by breaking them down to microtask level (Schmitz & Lykourantzou 2018). Such tasks are complex and would yield meaningless results if decomposed, because of the many interdependencies among knowledge domains that they entail, and the need to maintain the global context while working on them. One can think of writing a story (Kim et al. 2016), a news article, or defining a research methodology (Schmitz & Lykourantzou 2018). In juxtaposition to *microtasks*, this type of tasks are known as *macrotasks* (from the Greek word μακρός (makros) which means 'long, large'), and crowdsourcing research has just started to look into them (Haas et al. 2015, Cheng et al. 2015). Although ground-breaking, the aforementioned research in macrotask crowdsourcing has primarily used the term to contrast microtask crowdsourcing and in regards to the size of the task at hand, not its complexity and properties; like decomposition.

Macrotask crowdsourcing can make a more significant impact and to generate more value compared to microtask crowdsourcing, because it directly contributes to solving more challenging problems of both social and economic nature. Furthermore, it also requires salient, lifelong learning skills of the future such as creativity and critical thinking. By primarily focusing on microtasks, we are unnecessarily limiting and underestimating the crowd's potential.

Given the increasing interest of the research community and the industry on what can the crowds achieve, this book is a first effort to underpin this new type of crowd labor model that macrotask crowdsourcing represents and to collect works, of both theory and practice, around this subject that have started to emerge. In addition to researchers and practitioners interested in the evolution of crowdsourcing, it is our hope that this book will also prove useful for researchers and practitioners who are skeptical in regards to what they currently think what crowdsourcing is and what it can accomplish.

We initiate the book with a chapter that aims to properly define the terms *macrotask* and *macrotask crowdsourcing*. The chapter takes into account prior work and relevant theory, and looks deeper into the nature of the task, of worker skills and of crowd labor management, to provide a concrete basis upon which future researchers and practitioners can build upon. The rest of the book is divided into three parts, which together cover a wide range of macrotask crowdsourcing topics: *Coordination and Cooperation*, *The Role of AI and Experts*, and *Macrotasking for Social Good*.

Part I: Coordination and Cooperation

In this first part, the book examines the role of coordination and cooperation in the context of macrotasking. Coordination, in the context of complex work, is not an evident feat. Beyond issues of different time zones, languages, and cultures (issues that might anyway arise in microtasking) the multiple knowledge interdependencies and interactions required among the different workers create novel coordination challenges for macrotasks.

The Chap. 2 aims to advance our understanding on exactly this topic. More specifically, this chapter reviews several popular theories of coordination, examines the current approaches to crowd coordination in the HCI and CSCW literature, and identifies literature shortcomings. Based on these findings, the authors then proceed in proposing a research agenda and design propositions for each of the recommended theories of coordination, thus advancing our understanding of which crowd coordination mechanisms to select when complex macrotask work is involved.

A topic close to crowd coordination is crowd control. Crowdsourcing controls are mechanisms to align crowd workers' actions with predefined standards to achieve a set of goals and objectives set by the task requester. In ordinary microtasking, it is usually enough to address issues of control indirectly through financial incentives. In macrotasking, however, where the task is often performed within groups, more fine-grained behavior influencing control mechanisms are necessary to ensure a successful completion of the macrotask. In Chap. 3, the authors aim to develop a better comprehension of the controls appropriate for macrotask crowdsourcing. To accomplish this, they present and discuss the literature on control theory, identify a series of gaps, and put forth a research agenda to address these shortcomings. The proposed research agenda focuses on understanding how to design controls that are more suitable for macrotasking and the implications that such controls have for future crowdsourcing organizations.

This part of the book ends with an exploration of cooperation among crowd workers. Cooperation is an issue of less importance for microtasking, where workers usually perform tasks individually, but of increasing importance in macrotasking, where workers interact more often. In Chap. 4, the authors aim to leverage cooperation possibilities to improve the data quality of deployed macrotasks. The authors analyze three use cases from the domain of situated crowdsourcing, and use the results of this analysis to propose the design of a novel situated crowdsourcing platform that can effectively support cooperation without alienating solo workers.

Part II: The Role of AI and Experts

The second part of the book examines the role that Artificial Intelligence and Experts play in accomplishing macrotasks. As tasks become more complex, and in order to maintain their quality and scalability, advanced AI is becoming a necessity to efficiently distribute work among expert and nonexpert workers, as well as computational systems. Chapter 5 sheds light on exactly this topic. Using as an example, the macrotask of supporting scientific research at scale, the authors review the state-of-the-art in the intersection of crowdsourcing and AI, and outline how crowd computing research can inform the development of intelligent crowd-powered systems that can efficiently support macrotasking processes.

Selecting suitable workers has always been an important issue, ever since microtask crowdsourcing emerged. This selection is even more important in macrotasking, where the macrotask may require different types and granularities of expertise. In Chap. 6, the authors aim to ensure that the most appropriate workers will participate in the available tasks of a macrotask crowdsourcing marketplace. The authors base their work presenting two novel preselection mechanisms that have been shown to be effective in microtask crowdsourcing, and then proceed to discuss how these mechanisms can be used within macrotasks.

In the dawning age of macrotask crowdsourcing, should experts feel threatened? In the final chapter of this part of the book, the authors of the Chap. 7 present a highly reflective work of how digital technology could allow wider participation whilst preserving the core values of academia. Crucially, they address the question: Is academic resistance to crowdsourcing an elitist fear of the unwashed, or justifiable wariness of incipient poor scholarship?

Part III: Macrotasking for Social Good

As with every technology, macrotask crowdsourcing should eventually bring a positive development to future generations. In this part of the book, we present three chapters that showcase the potential broad benefits that macrotask crowdsourcing could bring to societal challenges.

Changing behaviors is a well-known challenge both widely acknowledged in HCI as well as other scientific fields. In Chap. 8, the authors aim to address this challenge by studying the effects of the content, mode, and style of motivational messages in the context of behavior change. To accomplish this, they use crowdsourcing for collecting a large amount of data to form an accessible database of motivational messages. The authors then report findings on unsupervised explorations of the emotional expressiveness and sound quality (signal-to-noise ratio, SNR) of the crowdsourced motivational speech.

Providing appropriate feedback is a crucial part of the learning process in educational setting. In Chap. 9, the authors aim to investigate how to compliment academic feedback with crowdsourced feedback. To accomplish this, they (1) investigate complimenting academic feedback with “real world feedback” during a course on mobile development, using HCI methods and (2) report the costs and benefits that both staff and students should be aware of, when planning to apply such methods.

Recent disasters due to climate change have been, rightfully so, prominently presented in popular media channels. In Chap. 10, the authors compare and contrast how different online communities employ crowdsourcing to aid disaster response efforts. To accomplish this, they first interview members from Humanitarian OpenStreetMap (HOT) and Public Lab mapping communities. Based on these

interviews, they employ OpenStreetMap Analytics and Social Network Analysis, and analyze community strategies and interface logistics involved in the work of both communities.

Eindhoven, The Netherlands
 Suzhou, China
 Utrecht, The Netherlands
 Eindhoven, The Netherlands

Vassillis-Javed Khan
 Konstantinos Papangelis
 Ioanna Lykourantzou
 Panos Markopoulos

References

- à Campo, S., Khan, V. J., Papangelis, K., & Markopoulos, P. (2018). Community heuristics for user interface evaluation of crowdsourcing platforms. *Future Generation Computer Systems*; Vol. 95, pp. 775–789. <https://doi.org/10.1016/j.future.2018.02.028>.
- Bernstein, M. S., Little, G., Miller, R. C., Hartmann, B., Ackerman, M. S., Karger, D. R., Crowell, D. & Panovich, K. (2010). Soylent: A word processor with a crowd inside. In *Proceedings of the 23rd Annual ACM Symposium on User Interface Software and Technology (UIST '10)* (pp. 313–322). ACM, New York, NY, USA.
- Cheng, J., Teevan, J., Iqbal, S. T., & Bernstein, M. S. (2015, April). Break it down: A comparison of macro-and microtasks. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (pp. 4061–4064). ACM.
- Haas, D., Ansel, J., Gu, L., & Marcus, A. (2015). Argonaut: Macrotask crowdsourcing for complex data processing. *Proceedings of the VLDB Endowment*, 8(12), 1642–1653.
- Kim, J., Sterman, S., Cohen, A. A. B., & Bernstein, M. S. (2017). Mechanical novel: Crowdsourcing complex work through reflection and revision. In *Proceedings of the 2017 ACM Conference on Computer-supported Cooperative Work and Social Computing* (pp. 233–245). New York, NY: ACM.
- Kittur, A., Chi, E. H., & Suh, B. (2008). Crowdsourcing user studies with mechanical turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '08)* (pp. 453–456). ACM, New York, NY, USA
- Lasecki, W., Miller, C., Sadilek, A., Abumoussa, A., Borrello, D., Kushalnagar, R., & Bigam, J. (2012). Real-time captioning by groups of non-experts. In *Proceedings of the 25th Annual ACM Symposium on User interface Software and Rechnology* (pp. 23–34). ACM.
- Luther, K., Pavel, A., Wu, W., Tolentino, J. L., Agrawala, M., Hartmann, B., & Dow, S. P. (2014). CrowdCrit: Crowdsourcing and aggregating visual design critique. In *Proceedings of the Companion Publication of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing* (pp. 21–24). ACM.
- Malone, T. W., Laubacher, R., & Dellarocas, C. (2010). The collective intelligence genome. *MIT Sloan Management Review*, 51(3), 21–31.
- Schmitz, H., & Lykourantzou, I. (2018). Online sequencing of non-decomposable macrotasks in expert crowdsourcing. *ACM Transactions on Social Computing*, 1(1), 1.
- Von Ahn, L., & Dabbish, L. (2004). Labeling images with a computer game. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 319–326). ACM.
- Von Ahn, L. (2008). Human computation. In *Proceedings of the 2008 IEEE 24th International Conference on Data Engineering* (pp. 1–2). IEEE Computer Society.

Contents

1	Macrotask Crowdsourcing: An Integrated Definition	1
	Ioanna Lykourantzou, Vassillis-Javed Khan, Konstantinos Papangelis and Panos Markopoulos	
Part I Coordination and Cooperation		
2	Crowdsourcing Coordination: A Review and Research Agenda for Crowdsourcing Coordination Used for Macro-tasks	17
	Sangmi Kim and Lionel P. Robert Jr.	
3	Crowdsourcing Controls: A Review and Research Agenda for Crowdsourcing Controls Used for Macro-tasks	45
	Lionel P. Robert Jr.	
4	Addressing Cooperation Issues in Situated Crowdsourcing	127
	Jorge Goncalves, Simo Hosio, Niels van Berkel and Simon Klakegg	
Part II The Role of AI and Experts		
5	Hybrid Machine-Crowd Interaction for Handling Complexity: Steps Toward a Scaffolding Design Framework	149
	António Correia, Shoaib Jameel, Hugo Paredes, Benjamim Fonseca and Daniel Schneider	
6	What You Sow, So Shall You Reap! Toward Preselection Mechanisms for Macrotask Crowdsourcing	163
	Ujwal Gadiraju and Mengdie Zhuang	
7	Crowdsourcing and Scholarly Culture: Understanding Expertise in an Age of Popularity	189
	Alan Dix, Rachel Cowgill, Christina Bashford, Simon McVeigh and Rupert Ridgewell	

Part III Macrotasking for Social Good

8 “You Can Do It!”—Crowdsourcing Motivational Speech and Text Messages 217
Roelof A. J. de Vries, Khiet P. Truong, Jaebok Kim and Vanessa Evers

9 Crowdsourcing Real-World Feedback for Human–Computer Interaction Education 233
Fernando Loizides, Kathryn Jones, Carina Girvan, Helene de Ribaupierre, Liam Turner, Ceri Bailey and Andy Lloyd

10 The Mapping Crowd: Macrotask Crowdsourcing in Disaster Response 253
Ned Prutzer

Chapter 1

Macrotask Crowdsourcing: An Integrated Definition



**Ioanna Lykourantzou, Vassillis-Javed Khan, Konstantinos Papangelis
and Panos Markopoulos**

Abstract The conceptual distinction between microtasks and macrotasks has been made relatively early on in the crowdsourcing literature. However, only recently a handful of research works has explored it explicitly. These works, for the most part, have focused on simply discussing macrotasks within the confines of their own work (e.g., in terms of creativity), without taking into account the multiple facets that working with such tasks involves. This has resulted in the term “macrotask” to be severely convoluted and largely meaning different things to different individuals. More importantly, it has resulted in disregarding macrotask crowdsourcing as a new labor model of its own right. To address this scholarly gap, in this paper we discuss macrotask crowdsourcing from a multitude of dimensions, namely the nature of the problem it can solve, the crowdworker skills it involves, and the work management structures it necessitates. In view of our analysis, we provide a first integrated definition of macrotask crowdsourcing.

1.1 Introduction

The distinction between microtasks and macrotasks was made relatively early on in the crowdsourcing literature. Grier (2013) emphasized the skills and expertise of workers when discussing macrotasks which he considers as “the professional form of crowdsourcing” and “freelancing on a global scale”, which happens in an open, public market contrary to microtasks, which are brief tasks that do not require advanced skills. Crowdsourcing platforms help manage the relationship between the requester

I. Lykourantzou (✉)

Department of Information and Computing Sciences, Utrecht University, Princetonplein 5, De
Uithof, 3584 CC Utrecht, The Netherlands
e-mail: i.lykourantzou@uu.nl

V.-J. Khan · P. Markopoulos

Eindhoven University of Technology, Eindhoven, The Netherlands

K. Papangelis

Xi’an Jiaotong-Liverpool University, Suzhou, China
e-mail: K.papangelis@xjtlu.edu.cn

© Springer Nature Switzerland AG 2019

V.-J. Khan et al. (eds.), *Macrotask Crowdsourcing*,

Human–Computer Interaction Series, https://doi.org/10.1007/978-3-030-12334-5_1

who owns the problem and the worker who will execute it, they handle payments, and support practical challenges such as verifying the time worked. Grier, like other authors after him, introduced macrotasks in juxtaposition to microtasks in terms of the magnitude of the task. These works go as far as to propose a checklist for defining a macrotask as follows: a macrotask is a task that can be carried out independently without support by the requester, which is simple to describe with clear criteria of completion, which has a clear and concrete deadline, and which requires special skills that the requester's organization does not possess. This practical and down-to-earth guidance helps get one on the way with macrotasking but does not shed much light into how macrotasking differs and why it needs to be addressed differently than microtasking.

One of the early investigations of task decomposition in crowdsourcing was presented in the case of video annotation (Vondrick et al. 2013). Video annotation is a canonical example of a crowdsourcing task where valuable results are obtained by combining small contributions by many crowdworkers. To assess the value of task decomposition Vondrick et al. (2013) compared annotating video for a single object per crowdworker which they considered as a microtask to annotating a video segment for a whole set of objects which they considered to be a macrotask. They noted how video annotation of a segment for all objects may cost more time but it allows the crowdworker to develop ownership of the result and deliver labels of higher quality. Furthermore, errors in coding specific objects are distributed over different segments and handled by different coworkers, while the effort a crowdworker invests to visually decode a scene is committed only once for all objects that need to be identified. Beyond video annotation, Machado et al. (2014) discuss crowdsourcing in the context of software development, where in line with Grier (2013) discussed above, they consider macrotasks as larger than microtasks and requiring specific knowledge from the crowdworker. They propose software testing as an example of a macrotask and discuss macrotasking practices by the Brazilian company Crowdtest or the American Utest.

Cheng et al. (2015) is the first (and to this point the only) empirical study that focuses explicitly on the trade-offs involved in decomposing macrotasks to microtasks. They examined task performance for three types of tasks, which included simple arithmetic, sorting text, and audio transcription. Their results suggest that decomposing macrotasks to smaller parts, may make the total task completion time longer but it enhances the task quality and makes work easier. The experiment and their whole discussion considers macro and microtasks as relative descriptions, the latter being a decomposition of the former. The macrotasks in their experiment are very simple, namely adding 10 numbers, sorting 7 lines of text or transcribing 30 seconds of audio. This helps test the decomposition decision very directly in the experiment, but does not help transposing the conclusions of this experiment to situations where leadership, creativity, initiative, coordination might be manifested, as it is often the case in what one might consider a more complex task in real life. Cheng et al. (2015), also considered how interruptions may affect the task completion time arguing that macrotasks are less resilient to interruptions. However, this result may indeed be very specific to the nature of the experimental tasks that they used, where task decomposition translates directly to lower demands on short term memory—which is

challenged during interruptions. Arguably decomposing macrotasks of much larger scale such as creating a logo, which might take minutes or hours rather than seconds, is not likely to produce similar gains.

Haas et al. (2015) identify quality control as one of the major challenges in setting up workflows involving macrotasking. They consider macrotasks as ones that cannot be easily decomposed, or where larger context (e.g., domain knowledge) or a significant initial investment of time is needed before workers can engage in task execution in order to develop a global context, e.g. when authoring a paper or a presentation. They point out that while crowdsourcing researchers have sought efficiency and quality gains in the algorithmic decomposition of tasks and synthesis of individual crowdworker microcontributions, there can be substantial benefits in recruiting task workers to perform macrotasks that last longer and which apply more flexible compensation schemes, combining some of the benefits of microtasks and traditional freelance work. Haas et al. (2015) introduce Argonaut, a framework for managing macrotask based workflows that addresses a major challenge for automating macrotask work, which is to ensure the quality of the work. The Argonaut framework profiles workers in terms of the work quality they deliver and their speed, and uses these profiles to sustain a hierarchy of roles (workers, reviewers, and top-tier reviewers). Workers are assigned suitable roles within the macrotask workflow and are promoted or demoted dynamically depending on task availability.

Li et al. (2016) consider macrotasks as those lasting several hours. They argue that workers are not easily motivated to carry out these, and that they are challenging to define/decompose. For this, they suggest that macrotasking is an important topic for future research.

Valentine et al. (2017) report on an approach for handling a specific class of macrotasks that are complex and open-ended, and which are difficult to crowdsource using microtasking because it is difficult to articulate, modularize, and prespecify the actions needed to achieve them. To do so, they propose ways to structure the crowd in “flash organizations” that involve defining formal structures such as roles, teams, and hierarchies that delineate responsibilities, interdependencies, and information flow without prespecifying all actions. Their approach is characterized by (a) a de-individualized role hierarchy (as can be found in organizations like movie crews, disaster response teams, or the army) where collaboration is based on workers’ knowledge of the roles rather than their knowledge of each other: (b) a continuous reconfiguration of the organization e.g., by changing roles or adding teams. Valentine et al. (2017) demonstrate the feasibility of their approach through three case studies concerning respectively: (1) creating an application for emergency medical technicians (EMTs) to report trauma injuries from an ambulance en route to the hospital designing, manufacturing, and playtesting a storytelling card game and an accompanying mobile application, and creating an enterprise web portal to administer client workshops.

Implementing such organizational structures in crowdsourcing in order to support macrotasks brings about challenges related to incentivizing workers. For example, personal preferences or biases may color assessments of solution quality. Xie and Lui (2018) propose an optimization approach for incentivizing workers to provide

high-quality contributions and empirically evaluate the effectiveness and efficiency of their approach.

1.2 On the Nature of the Problem

To understand the reasons that may necessitate a shift from microtasking to macro-tasking, one must first understand the problems that each crowdsourcing model can and cannot solve. Drawing from organizational management literature, below we classify crowdsourcing models according to the problem attributes that each can solve (Fig. 1.1).

Knowledge problems can be categorized based on three attributes: complexity, decomposability, and structure (Nickerson and Zenger 2004; Huang and Holden 2016). **Complexity** refers to the number of knowledge domains that are relevant to the problem, and the strength of their interactions. Simple problems tend to involve few knowledge domains, with a low degree of domain interdependency. More complex problems involve a large number of knowledge domains, which share a strong degree of domain interaction. **Decomposability** measures whether the problem can be divided into subproblems, and the granularity that this division can reach. Decomposable problems can be broken down to separate subproblems, each drawing from distinct knowledge sets, which can be solved independently and with little communication or collaboration among problem solvers. Non-decomposable problems on

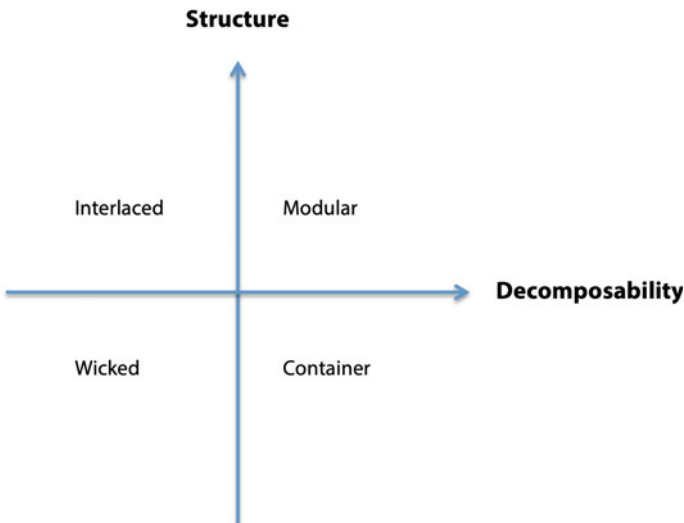


Fig. 1.1 The macrotask dimension space. To draw this diagram we assume all macrotasks are complex. Then we have a cartesian space of them with the dimensions of structure and decomposability. This space characterizes four types of macrotasks: interlaced, modular, wicked and container

the other hand, are impractical or even impossible to subdivide into separate sub-problems, because the interdependencies among their knowledge domains are too extensive. For such problems, if a solution is to be found, this needs to be an overall solution, which enables problem solvers to maintain the global problem context. **Structure** is the degree to which one can determine all the knowledge domains relevant to the problem, the expertise needed to solve it, and the interrelations between the identified domains. Well-structured problems consist of a clear set of relevant knowledge domains. The boundaries and interactions among these domains can be easily understood, and there are explicit and widely accepted approaches to solve the problem. On the other hand, ill-structured problems are those where the relevant knowledge domains, necessary to solve the problem, are not evident, the boundaries among these domains are ambiguous and their in-between interactions are very poorly understood. Conversely, consensus approaches may not be optimal; rather these problems often benefit from “spontaneous” disruptive innovations, which often challenge scientific and industrial status quos and offer new ways of interpreting the problem and its solution.

This classification enables us to position existing and future crowdsourcing models with respect to the problems that they can solve, and the problems for which they are not suitable.

Tasks related to data such as: categorization, curation, or enrichment (Kittur et al. 2008; Musthag and Ganesan 2013) tackle problems that are simple, well-structured, and decomposable. The bulk of tasks in most commercial crowdsourcing platforms are of that sort.

1.2.1 Macrotask Type 1 (Modular): Well-Structured, High-Decomposability Problems

The first type of macrotasks is meant to solve problems that are, *decomposable*, and *well-structured*. These form the majority of complex problems that current crowdsourcing literature and applications focus on, and understandably so, since these problems can be addressed using a “*divide and conquer approach*”. The problem is first broken down to smaller, distinct work units, i.e., at microtask level. Then, the distinct microtasks are assigned in parallel to multiple workers, and finally they are recomposed to a final output by combining the separate smaller subtasks.

The difference with what we might call “vanilla” microtasking is that, because of the problem complexity, the way of breaking down the problem to microtasks is not evident and may require the *involvement of experts, who design tailor-made workflows for the crowd to follow*. These experts in collaboration with the task requester, often determine how the macrotask should be decomposed into smaller chunks, and how to recompose these once completed. Because of the involvement of experts the decomposition of microtask level can be costly (Kim et al. 2014; Chan et al. 2016). Nevertheless, once the workflow has been designed, it can be very effective

(Teevan et al. 2016). That being said, this approach suffers from non-generalization. Because the workflows are usually tailored to the very specific problem, they cannot be generalized easily to handle other problem instances.

The resulting microtasks may not be homogeneous in terms of size, or skill requirement.

Examples of macrotask type 1 include: taxonomy creation (Chilton et al. 2013), itinerary planning (Zhang et al. 2012), editing and correcting a document (Bernstein et al. 2010), or aggregating multiple word or sentence-level translations to form a larger corpus (Ambati et al. 2012; Zaidan and Callison-Burch 2011).

1.2.2 Macrotask Type 2 (Interlaced): Well-Structured, Low-Decomposability Problems

The second type of macrotasks aims to tackle problems that are well-structured but are non-decomposable. In general, these are problems often found at the beginning of creative projects (e.g., when the broad objectives and solution criteria need to be set) and are, for the most part, only processed manually, even if the rest of the project can be broken down into subtasks and potentially crowdsourced (Sieg et al. 2010). These problems can be solved through a “continuity of useful action” (Altshuller 2005) where each consecutive contributor maintains the global context and full semantic overview of the problem while iteratively refining it until an acceptable solution is found.

Examples of type 2 macrotasks would be: defining a research methodology or formulating an R&D approach.

1.2.3 Macrotask Type 3 (Wicked): Ill-Structured, Low-Decomposability Problems

The third type of macrotask problems are the so-called “wicked problems” or “holy grail” problems. These are ill-structured tasks, for which the interactions among the relevant knowledge domains (or even the exact required knowledge domains themselves), are not well understood, and the requirements are incomplete, contradictory, and in some cases ever-changing. Wicked problems, in a crowdsourcing context, tend to be handled through innovation idea contests (Majchrzak and Malhotra 2013), where the purpose is to collect as many ideas as possible in search for the few breakthrough ideas, rather than an iterative idea development. There has been limited research on how to process and tackle wicked problems through crowdsourcing. Evidence illustrates that using a sequential process could lead to problems such as fixation with one solution (Jansson and Smith 1991) or solution confounding (Little et al. 2010). However, further research is necessary to shed light on the issue.

An example of a type 3 macrotask is end-to-end innovation production.

1.2.4 Macrotask Type 4 (Container): Ill-Structured, High-Decomposability Problems

The final macrotask type aims to tackle problems that are ill-structured and highly decomposable. Although such problems are not directly addressed in the literature, one could conceptually identify them based on the structure/decomposability matrix that organizational research suggests. Such problems could be those for which the required expertise cannot be determined automatically a priori, but it can be determined with the help of an expert or team of experts. For example, in a crowdsourcing context, such a problem is the coordination of a team of crowd workers. Very recent literature (Wood et al. 2019) has indeed touched upon this phenomenon, reporting that high-reputation crowd workers delegate complex work to other crowd workers or other workers from their social circles. They also often explain the tasks and train (in the form of instructions) their delegates on how to accomplish the (part of) complex work. This method of understanding the ill-structured problem, and then decomposing and delegating it based on experience, could be a precursor of more complex workflows that are needed to handle this type of tasks. Future work is required to research such problems in more detail, and understand which crowdsourcing workflows can be designed to address them.

1.3 On the Nature of Skills

Few works in existing microtask crowdsourcing literature focus on workers skills. Although very recent works in the area do try to understand better the needs of the crowdworkers, for example by examining their working conditions or the context they find themselves into (Gray et al. 2016; Irani and Silberman 2013; Martin et al. 2014), these works do not examine which skills a worker has or needs to have. This research gap may be partially attributed to the fact that, apart from language (e.g., English) skills and general perception skills, workers in microtask crowdsourcing are usually not required to have very specialized skills to perform their work. Consequently, microtasking platforms also usually store only worker demographics and the percentage of tasks the worker has successfully completed (number of HITs, Levels, or other name depending on the platform). Microtasking platforms do not usually store other worker skills (Ho and Vaughan 2012). In case requesters need workers to have a specialized skill, they mention it in an open field, which workers fill in based on self-assessment. Self-assessment may be biased and its validity as a metric of skill quality is low since not all workers have the same perception of their skills. Less often, requesters may develop a tailor-made test, prior to the actual

microtask, to test specialized worker skills. This practice however is costly, and not generalizable.

In addition, microtasking usually relies on *skill homogeneity*: the problem is decomposed to microtasks that all require the same type of nonexpert skill. Consequently, currently not a lot of works in existing crowdsourcing literature analyze the spectrum of worker skills across a variety of possible problems that they could solve. The only works that usually assume a variety of different skills are based on simulations, either across different domains of the same level (Basu Roy et al. 2015), or even across hierarchical skills levels (Mavridis et al. 2016).

Macrotasking on the other hand is innately linked with skill diversity, and more fine-grained skill types, including expert and twenty-first-century skills, as well as valid skill identification and evaluation mechanisms. Examples of higher order cognitive and twenty-first-century skills that macrotask workers might need include: creativity, curiosity and imagination, critical thinking and problem-solving (Creative and Cultural Skills 2017), effective oral and written communication skills, information analysis ability, agility, adaptability and the capacity to learn new knowledge fast, collaboration ability, communication skills, taking initiative, leadership and people management skills (Wagner 2014). Expert skills can be obtained by direct training and “learning by doing”, and naturally include the whole spectrum of today’s and tomorrow’s expertise, with some prominent examples being coding, graphic design skills, research methodology skills, business marketing and communications, etc.

Although microtask crowdsourcing practice tends to consider workers as an endless, homogeneous and replaceable mass, the truth is that complex skills and crowd workers who possess them are inevitably expected to be less frequent. Therefore, for macrotask crowdsourcing, it is important to ensure the following:

- **Skill structure and assessment.** Develop mechanisms to assess macrotasking skills with validity, and in a scalable manner (Ipeirotis and Gabrilovich 2014), drawing from a wide range of approaches (from computerized to peer assessment), as well as the skill assessment scientific domain.
- **Develop training opportunities.** Workers who are not at the right skill level should not be excluded at face value. Rather, macrotasking platforms should support worker skill development, by offering training opportunities and scaffolded learning.
- **Access to skill data and skill data sharing.** Provide workers with expert skills with an access to and ownership of their skill data, and the opportunity to share them across platforms. This approach is not only in line with latest data management ethics (see the recent EU GDPR rules, see Voigt and Bussche 2017), but it is also expected to give workers a sense of control, the ability to indicate their skill pertinency, and promote workers mobility and platform cross-fertilization.

1.4 On the Nature of Management

When referring to crowdsourcing, scalability is the key. Unlike traditional management settings, where the human manager needs to organize the work of a few people (up to the level of dozens), the scale of crowdsourcing necessitates automation. For this reason, recent works have focused on algorithm-based human resource allocation in crowdsourcing settings, from two perspectives. From the mathematical optimization perspective, such algorithms assume a large pool of worker profiles (skills, availability, etc.) and a large pool of tasks with certain characteristics (e.g., knowledge domain), and constraints (deadline, budget, etc.). In this setting, the objective of the algorithms is to match each task with one or more workers, to accomplish the task optimally (e.g., in terms of quality) with the given constraints (e.g., Basu Roy et al. 2015; Goel et al. 2014; Schmitz and Lykourantzou 2018). From an organizational perspective, viewing crowds as organizations, algorithms coordinate the automated hiring of workers for different roles, and computationally structure their activities around complex workflows (Retelny et al. 2014; Kim et al. 2014; Valentine et al. 2017). Other types of algorithms, focusing more on teamwork, computationally rotate workers in different team combinations, to mix their viewpoints and ideas (Salehi and Bernstein 2018).

The problem with existing crowd management algorithms, is that they tend to **micro-manage the workers**, by assigning them directly on a specific task or team. Existing algorithms also tend to focus on computational efficiency and optimization. This approach is appropriate for microtasking, but it has drawbacks when it comes to macrotasks, as it can stifle creativity and initiative-taking, as indicated by recent research in management sciences (Lawler and Worley 2006) and crowdsourcing (Retelny et al. 2017). Future research is therefore needed to explore flexible algorithms that avoid micromanaging the workers, and explore ways to empower them.

Furthermore on crowd management, current crowdsourcing platforms have usually two management levels, i.e., the requester and the worker. Very recent works, indicate that new, multilevel ways of organization, such as re-outsourcing (Wood et al. 2019) and subcontracting (Morris et al. 2017), and Upwork's agency structures are emerging. Although the above works are applied on microtasking and freelance work, the multilevel management approach that they propose could be especially beneficial for the needs of microtasking (see microtask types 2, 3, and 4 above). Future research could explore this dimension.

A final note on crowd management is incentives engineering. Current microtasking crowdsourcing primarily relies on monetary rewards. Prior research in this domain has shown that higher payment indeed leads to faster completion time of the microtasks, but not necessarily to higher quality (Mason and Watts 2009). Initial research shows that purely extrinsic motivators, such as money, are not enough (Zheng et al. 2011). Microtasking, which often involves open-ended and innovation-oriented work, and which for this reason relies on workers' creativity and expertise,

needs to find the right balance between extrinsic and intrinsic incentives. Earlier studies have offered “implications for the design of mobile workforce services, including future services that do not necessarily rely on monetary compensation” (Teodoro et al. 2014). For this reason, further work is needed to explore which intrinsic incentives platforms could offer to motivate quality macrotask work; examples might include: providing work feedback, and scaffolding workers’ career growth (Edmondson et al. 2001). To ensure that this research will have practical impact, crowdsourcing platforms need to raise awareness and educate requesters about the importance of offering such incentives and support them in the process of doing so.

1.5 Macrotask Crowdsourcing Definition

Taking into account the aforementioned dimensions, on the nature of the task, the skills of the workers, and the management principles, we provide below a first integrated definition of macrotask crowdsourcing:

Macrotask crowdsourcing refers to crowdsourcing that is designed to handle complex work of different degrees of structure and decomposability, assumes varying levels of (expert) knowledge over one or more domains, requires a range of 21st century skills, benefits from worker communication, collaboration, and training, and incorporates flexible work management processes that potentially involve the workers.

1.6 Conclusion

In this chapter we discuss macrotask crowdsourcing in terms of three dimensions: (i) the complex *problems* this labor model can solve, (ii) the worker *skills* it requires and (iii) the *management structures* it benefits from. In regards to the first dimension, we define four types of macrotasks—modular, interlaced, wicked, and container. Each type can solve a different problem, based on two problem axes: decomposability and structure. Regarding the second dimension, we touched upon the worker skills required for macrotask crowdsourcing, emphasizing the need for skill diversity, fine-grained skill types, expert and twenty-first-century skills, as well as for skill development and evaluation mechanisms. Finally, in regards to the third dimension, we discussed the work management structures that are appropriate for this new type of work, highlighting the need to avoid micromanaging the workers but rather providing them with more initiative and actively involving them in the management of their work. We conclude this chapter with a definition, for the first time, of macrotask crowdsourcing. Our aim in providing this definition is to assist future researchers to better position their work, and inspire future developments in this expanding field.

References

- Altshuller, G. (2005). *The innovation algorithm: TRIZ, systematic innovation and technical creativity*. Worcester, MA: Technical Innovation Center.
- Ambati, V., Vogel, S., & Carbonell, J. (2012). Collaborative workflow for crowdsourcing translation. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work (CSCW '12)* (pp. 1191–1194). ACM, New York, NY, USA.
- Basu Roy, S., Lykourantzou, I., Thirumuruganathan, S., Amer-Yahia, S., & Das, G. (2015). Task assignment optimization in knowledge-intensive crowdsourcing. *The VLDB Journal—The International Journal on Very Large Data Bases*, 24(4), 467–491.
- Bernstein, M. S., Little, G., Miller, R. C., Hartmann, B., Ackerman, M. S., Karger, D. R., Crowell, D., & Panovich, K. (2010). Soylent: A word processor with a crowd inside. In *Proceedings of the 23rd Annual ACM Symposium on User Interface Software and Technology (UIST '10)* (pp. 313–322). ACM, New York, NY, USA.
- Chan, J., Dang, S., & Dow, S. P. (2016). Improving crowd innovation with expert facilitation. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing (CSCW '16)* (pp. 1223–1235). ACM, New York, NY, USA.
- Cheng, J., Teevan, J., Iqbal, S. T., & Bernstein, M. S. (2015, April). Break it down: A comparison of macro-and microtasks. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (pp. 4061–4064). ACM.
- Chilton, L. B., Little, G., Edge, D., Weld, D. S., & Landay, J. A. (2013). Cascade: Crowdsourcing taxonomy creation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)* (pp. 1999–2008). ACM, New York, NY, USA.
- Creative and Cultural Skills. (2017). Building a creative nation: Current and future skills need. https://ccskills.org.uk/downloads/Building_a_Creative_Nation_-_Current_and_Future_Skills_Needs.pdf.
- Edmondson, A. C., Bohmer, R. M., & Pisano, G. P. (2001). Disrupted routines: Team learning and new technology implementation in hospitals. *Administrative Science Quarterly*, 46(4), 685–716.
- Goel, G., Nikzad, A., & Singla, A. (2014). Allocating tasks to workers with matching constraints: Truthful mechanisms for crowdsourcing markets. In *Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion (WWW Companion '14)* (pp. 279–280). International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland.
- Gray, M. L., Suri, S., Ali, S. S., & Kulkarni, D. (2016, February). The crowd is a collaborative network. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing* (pp. 134–147). ACM.
- Grier, D. A. (2013). *Crowdsourcing for dummies*. Wiley.
- Haas, D., Ansel, J., Gu, L., & Marcus, A. (2015). Argonaut: Macrotask crowdsourcing for complex data processing. *Proceedings of the VLDB Endowment*, 8(12), 1642–1653.
- Ho, C. J., & Vaughan, J. W. (2012, July). Online task assignment in crowdsourcing markets. In *Twenty-sixth AAAI Conference on Artificial Intelligence*.
- Huang, S., & Holden, D. (2016). The R&D boundaries of the firm: A problem solving perspective. *International Journal of the Economics of Business*, 23(3), 287–317.
- Ipeirotis, P. G., & Gabrilovich, E. (2014, April). Quizz: Targeted crowdsourcing with a billion (potential) users. In *Proceedings of the 23rd International Conference on World Wide Web* (pp. 143–154). ACM.
- Irani, L. C., & Silberman, M. (2013, April). Turkopticon: Interrupting worker invisibility in amazon mechanical turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 611–620). ACM.
- Jansson, D. G., & Smith, S. M. (1991). Design fixation. *Design Studies*, 12(1), 3–11.
- Kim, J., Cheng, J., Bernstein, & M.S. (2014) Ensemble: Exploring complementary strengths of leaders and crowds in creative collaboration. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing* (pp. 745–755). CSCW '14,

- ACM, New York, NY, USA. <https://doi.org/10.1145/2531602.2531638>, <http://doi.acm.org/10.1145/2531602.2531638>.
- Kittur, A., Chi, E. H., & Suh, B. (2008). Crowdsourcing user studies with mechanical turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '08)* (pp. 453–456). ACM, New York, NY, USA
- Lawler, E. E., & Worley, C. G. (2006). Designing organizations that are built to change. *MIT Sloan Management Review*, 48(1), 19–23.
- Li, G., Wang, J., Zheng, Y., & Franklin, M. J. (2016). Crowdsourced data management: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 28(9), 2296–2319.
- Little, G. Chilton, L. B., Goldman, M., & Miller, R. C. (2010). Exploring Iterative and parallel human computation processes. In *Proceedings of the ACM SIGKDD Workshop on Human Computation (HCOMP '10)*. ACM, New York, NY, USA, 68–76.
- Machado, L., Pereira, G., Prikladnicki, R., Carmel, E., & de Souza, C. R. (2014, November). Crowdsourcing in the Brazilian IT industry: What we know and what we don't know. In *Proceedings of the 1st International Workshop on Crowd-based Software Development Methods and Technologies* (pp. 7–12). ACM.
- Martin, D., Hanrahan, B. V., O'Neill, J., & Gupta, N. (2014, February). Being a turker. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing* (pp. 224–235). ACM.
- Mason, W., & Watts, D. J. (2009, June). Financial incentives and the performance of crowds. In *Proceedings of the ACM SIGKDD Workshop on Human Computation* (pp. 77–85). ACM.
- Mavridis, P., Gross-Amblard, D., & Miklós, Z. (2016, April). Using hierarchical skills for optimized task assignment in knowledge-intensive crowdsourcing. In *Proceedings of the 25th International Conference on World Wide Web* (pp. 843–853). International World Wide Web Conferences Steering Committee.
- Majchrzak, A., & Malhotra, A. (2013). Towards an information systems perspective and research agenda on crowdsourcing for innovation. *The Journal of Strategic Information Systems*, 22(4), 257–268.
- Morris, M. R., Bigham, J. P., Brewer, R., Bragg, J., Kulkarni, A., Li, J., & Savage, S. (2017, May). Subcontracting microwork. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (pp. 1867–1876). ACM.
- Musthag, M., & Ganesan, D. (2013). Labor dynamics in a mobile micro-task market. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)* (pp. 641–650). ACM, New York, NY, USA.
- Nickerson, J. A., & Zenger, T. R. (2004). A knowledge-based theory of the firm—the problem-solving perspective. *Organization Science*, 15(6), 617–632.
- Retelny, D., Robaszekiewicz, S., To, A., Lasecki, W. S., Patel, J., Rahmati, N., ... & Bernstein, M. S. (2014, October). Expert crowdsourcing with flash teams. In *Proceedings of the 27th annual ACM Symposium on User Interface Software and Technology* (pp. 75–85). ACM.
- Retelny, D., Bernstein, M. S., & Valentine, M. A. (2017). No workflow can ever be enough: How crowdsourcing workflows constrain complex work. In *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW), 89.
- Salehi, N., & Bernstein, M. S. (2018). Hive: Collective design through network rotation. In *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW), 151.
- Schmitz, H., & Lykourantzou, I. (2018). Online sequencing of non-decomposable macrotasks in expert crowdsourcing. *ACM Transactions on Social Computing*, 1(1), 1.
- Sieg, J. H., Wallin, M. W., & Von Krogh, G. (2010). Managerial challenges in open innovation: a study of innovation intermediation in the chemical industry. *R&D Management*, 40(3), 281–291.
- Teevan, J., Iqbal, S. T., Cai, C. J., Bigham, J. P., Bernstein, M. S., & Gerber, E. M. (2016). Productivity decomposed: Getting big things done with microtasks. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '16)* (pp. 3500–3507). ACM, New York, NY, USA.

- Teodoro, R., Ozturk, P., Naaman, M., Mason, W., & Lindqvist, J. (2014, February). The motivations and experiences of the on-demand mobile workforce. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing* (pp. 236–247). ACM.
- Valentine, M. A., Retelny, D., To, A., Rahmati, N., Doshi, T., & Bernstein, M. S. (2017, May). Flash organizations: Crowdsourcing complex work by structuring crowds as organizations. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (pp. 3523–3537). ACM.
- Voigt, P., & Von dem Bussche, A. (2017). The EU General Data Protection Regulation (GDPR). In *A practical guide*, 1st edn. Cham: Springer.
- Vondrick, C., Patterson, D., & Ramanan, D. (2013). Efficiently scaling up crowdsourced video annotation. *International Journal of Computer Vision*, 101(1), 184–204.
- Wagner, T. (2014). *The global achievement gap: Why even our best schools don't teach the new survival skills our children need—and what we can do about it*. Basic Books.
- Wood, A. J., Graham, M., Lehdonvirta, V., & Hjorth, I. (2019). Networked but commodified: The (Dis) embeddedness of digital labour in the gig economy. *Sociology*, 0038038519828906.
- Xie, H., & Lui, J. C. (2018). Incentive mechanism and rating system design for crowdsourcing systems: Analysis, tradeoffs and inference. *IEEE Transactions on Services Computing*, 11(1), 90–102.
- Zheng, H., Li, D., & Hou, W. (2011). Task design, motivation, and participation in crowdsourcing contests. *International Journal of Electronic Commerce*, 15(4), 57–88.
- Zhang, H., Law, E., Miller, R., Gajos, K., Parkes, D., & Horvitz, E. (2012). Human computation tasks with global constraints. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)* (pp. 217–226). ACM, New York, NY, USA.
- Zaidan, O. F., & Callison-Burch, C. (2011). Crowdsourcing translation: Professional quality from non-professionals. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1 (HLT '11)* (pp. 1220–1229). Association for Computational Linguistics, Stroudsburg, PA, USA.

Part I
Coordination and Cooperation

Chapter 2

Crowdsourcing Coordination: A Review and Research Agenda for Crowdsourcing Coordination Used for Macro-tasks



Sangmi Kim and Lionel P. Robert Jr.

Abstract Crowdsourcing has become a widely accepted approach to leveraging the skills and expertise of others to accomplish work. Despite the potential of crowdsourcing to tackle complex problems, it has often been used to address simple micro-tasks. To tackle more complex macro-tasks, more attention is needed to better comprehend crowd coordination. Crowd coordination is defined as the synchronization of crowd workers in an attempt to direct and align their efforts in pursuit of a shared goal. The goal of this chapter is to advance our understanding of crowd coordination to tackle complex macro-tasks. To accomplish this, we have three objectives. First, we review popular theories of coordination. Second, we examine the current approaches to crowd coordination in the HCI and CSCW literature. Finally, the chapter identifies shortcomings in the literature and proposes a research agenda directed at advancing our understanding of crowd coordination needed to address complex macro-tasks.

2.1 Introduction

Crowdsourcing has become a widely accepted approach to leveraging the skills and expertise of others to accomplish work (Robert and Romero 2015, 2017). Crowdsourcing has many definitions but was first defined by Jeff Howe as the outsourcing of work to a crowd (Howe 2006). Typical modern definitions of crowdsourcing involve two attributes: (1) a crowd, or group of people, and (2) online work. Crowdsourcing platforms such as Mechanical Turk (<http://www.mturk.com>) and CrowdFlower (<http://www.crowdfower.com>) attract large groups of people who can work online via these digital platforms. These platforms and the people who work on them (i.e., crowd workers) provide access to a wealth of knowledge and expertise that can be leveraged to tackle complex problems.

S. Kim (✉) · L. P. Robert Jr.
University of Michigan School of Information, Ann Arbor, USA
e-mail: sangmik@umich.edu

L. P. Robert Jr.
e-mail: lprobert@umich.edu

© Springer Nature Switzerland AG 2019
V.-J. Khan et al. (eds.), *Macrotask Crowdsourcing*,
Human-Computer Interaction Series, https://doi.org/10.1007/978-3-030-12334-5_2

Despite the potential of crowdsourcing to tackle complex problems, it has often been used to address rather simple micro-tasks. Micro-tasks are standalone simple tasks that do not require the coordination of work among individuals (Schmitz and Lykourantzou 2018). To tackle more complex problems, crowdsourcing must address macro-tasking. Macro-tasking can be described as complex crowd work that is sometimes but not always decomposable to micro-tasks (Schmitz and Lykourantzou 2018). Crowdsourcing macro-tasks is more challenging than crowdsourcing micro-tasks. Macro-tasking requires work processes needed to tackle complex problem-solving involving activities such as the generation and integration of diverse ideas along with group decision-making. Macro-tasking also requires crowd workers to coordinate in order to both divide their labor and aggregate the outputs of their labor.

In the human–computer interaction/computer-supported cooperative work (HCI/CSCW) fields, crowd coordination is typically handled by the requestor and results in micro-tasking. Requestors divide and assign work prior to any crowd involvement and in many cases the work is never aggregated. Unfortunately, this approach to crowd coordination limits the potential of crowds to solve complex problems and reach their full potential.

Consider the following scenario: An organization wants to use crowdsourcing to identify its next new product. The organization puts forth a call to the public for new ideas and gives a specific deadline. The organization receives many great ideas and asks the crowd to vote on the best idea for a new product. The votes are tallied and the winner is announced. This approach to crowdsourcing is oriented toward micro-tasking. The work process is reasonably well formulated and easy to understand by all crowd workers. Although the outcome might not be predictable, the work process is very predictable. The crowdsourcing tasks require little interaction or dependence among crowd workers, so coordination is of little importance.

Now consider a different scenario: An organization wants to crowdsource the development of the marketing plan for this new product. Because there are many ways to accomplish this task, the work is not easily nor reasonably well formulated. Both the work process and the outcome are not as predictable as in the last scenario. Because the crowd is expected to produce one marketing plan, the crowd workers must decide how the work is to be divided and how or whether the work needs to be aggregated. To accomplish this task, crowd workers need to work together. This approach to crowdsourcing is oriented toward macro-tasking and requires interaction and greater dependence among crowd workers; therefore, coordination is of the utmost importance. Clearly, to fully leverage crowdsourcing, more work is needed on coordinating the crowdsourcing of macro-tasks.

There are many definitions of coordination (Robert 2016). For the sake of clarity, this chapter defines coordination generally as:

The synchronization of individuals in an attempt to direct and align their efforts in pursuit of a shared goal.

And crowd coordination specifically as:

The synchronization of crowd workers in an attempt to direct and align their efforts in pursuit of a shared goal.

The goal of this chapter is to advance our understanding of macro-tasking in crowdsourcing by addressing issues related to coordination. To accomplish this, we have three objectives. First, we review popular and recent theories of coordination across organizational and computer science. Specifically, we present and discuss transactive memory systems (TMS), coordination theory, role-based coordination, relational coordination, stigmergic coordination, and an integrative model of coordination. Second, we examine the HCI and CSCW studies on coordination in macro-tasking and categorize these approaches into one or more of the previously presented theories of coordination. Although prior studies on coordination in crowdsourcing have focused primarily on micro-tasking, attention is shifting toward macro-tasking, as seen by a small but fast-growing set of HCI/CSCW articles on the topic. Last, we propose a research agenda based on the review of coordination theories and prior HCI and CSCW work on coordination in macro-tasking.

2.2 Background

2.2.1 *Coordination in Micro-tasking Versus Macro-tasking in Crowdsourcing*

The first question one might ask is this: What makes coordinating macro-tasks so different from coordinating micro-tasks? Macro-tasks require much more coordination among workers than micro-tasks, for several reasons. Many micro-tasks are independent individual decomposed tasks assigned to individuals. Standalone independent micro-tasks require little or no coordination among crowd members. However, in many cases, macro-tasks cannot be decomposed to the level of a single individual and require more than one person to perform the work. The interdependent nature of macro-tasking requires coordination among crowd workers. In addition, macro-tasks that can be decomposed are likely to be decomposed by the crowd and not the requestor. Both the decomposition of macro-tasks and the eventual aggregation of micro-tasks require coordination among crowd members.

2.2.2 *Theories of Coordination*

2.2.2.1 **Transactive Memory System**

What is it? A transactive memory system (TMS) is a way of coordinating work that relies on members of a collective to know who knows what in that collective. This is accomplished in part by sharing or dividing the cognitive labor across the collective (Brandon and Hollingshead 2004; Wegner 1987). Research linking TMS to better coordination and ultimately performance has been conducted across a wide

and diverse set of fields including information systems, organizational behavior, psychology, and communications (Ren and Argote 2011). More specifically, the coordination benefits of TMS have led to emergent and adaptive team behaviors, allowing for effective and implicit communication (Marques-Quinteiro et al. 2013). TMS has proved to be an invaluable approach to team coordination.

How does it work? TMS effectiveness relies on five key elements. First, each member of the collective should hold unique specialized knowledge. Second, members of the collective should share a cognitive map of the distribution of this specialized knowledge across the team. Three, task responsibilities should be assigned to each member of the collective based on their specialized knowledge (Brandon and Hollingshead 2004; Moreland 1999). Four, members of the collective should trust that each member is competent in his or her knowledge domain and assigned task responsibilities (Austin 2003; Lewis 2003). If members of the collective do not trust one another they will be less likely to rely on one another's expertise. Five, members of the collective must communicate with one another to leverage each person's expertise (Choi et al. 2010). Communication allows for the sharing of specialized knowledge, which is essential for leveraging expertise across the collective.

Transaction Memory System Key Elements

- Specialized knowledge among members
- Shared cognitive map of specialized knowledge
- Task responsibilities based on specialized knowledge
- Members who trust one another's specialized knowledge
- Members who share their specialized knowledge

Potential Benefits for Macro-tasking: TMS allows for coordination among crowd workers through implicit communication. This reduces the overhead associated with explicit communication. TMS can also be used to organize and assign tasks. As new work requirements emerge, they are automatically assigned to crowd workers based on their knowledge specialization.

Potential Drawbacks for Macro-tasking: Crowds should have either a shared work history to develop a TMS or some way to communicate who knows what in a crowd. Developing a TMS can take time that crowd workers may not have. Platforms can be designed to communicate who knows what in a crowd. But it could be problematic for existing crowd workers to keep track of who knows what with regard to departing and incoming members.

2.2.2.2 Coordination Theory

What is it? Coordination theory is one of the most popular approaches to understanding coordination (Crowston et al. 2006). The theory defines coordination as the management of "dependencies between activities" (Malone and Crowston 1994, p. 90). One of the distinctive applications of the coordination theory is the use of

coordination mechanisms that are based on the type of dependencies among tasks for designing collaborative systems (e.g., Andres and Zmud 2002; Strode et al. 2012). Malone and Crowston (1994) introduced ways to analyze coordination in terms of actors, interdependent tasks that are performed by the actors, and resources that are required for completing those tasks. Based on their analysis, coordination problems that arise from the dependencies among tasks, actors, and resources are identified and solved by deploying appropriate coordination mechanisms.

How does it work? Several aspects of coordination theory make it distinct from other theories of coordination. First, it draws attention to the dependencies among tasks rather than among individuals or organizational units (Crowston et al. 2006). Instead of understanding coordination in terms of how people who perform the task relate to one another, this theory views coordination in terms of how one task is related to another task. Second, it identifies and categorizes types of dependencies among activities. This provides clarity as to possible implications associated with specific dependencies. Finally, this theory allows for the modeling of coordination to make it easier to understand the effects of assignments and reassignments of activities needed to complete tasks (Crowston 1994). This allows people to understand the implications of adding or removing members of the collective relative to that change's impact on coordination. However, recent work highlighted the limitations of coordination theory for coordinating crowd work (Retelny et al. 2017).

Coordination Theory Key Elements

- Identify tasks
- Identify and categorize dependencies among tasks
- Employ appropriate mechanism per dependency type

Potential Benefits for Macro-tasking: Coordination theory allows for the identification and removal of potential barriers to accomplishing crowd work. The workflow plans derived from coordination theory not only provide guidance needed to accomplish work but also a shared communication medium to facilitate a common understanding among crowd workers.

Potential Drawbacks for Macro-tasking: Coordination theory relies heavily on a person or group to pre-plan the work, which is less useful when task requirements are not known or task requirements are emergent and change over time. For example, at least one study has found evidence of this limitation as it relates to crowdsourcing complex adaptive work (see Retelny et al. 2017).

2.2.2.3 Role-Based Coordination Theory

What is it? Role-based coordination relies on roles or a set of expectations associated with a position to organize and perform work (Bechky 2006). Roles constitute both expected activities and their associated responsibilities. Roles have long been viewed in organizations as the basic unit of coordination (Okhuysen and Bechky 2009).

Role-based coordination does not rely on specific individuals, which has proved in some cases to be effective for complex and interdependent crowd work with transit membership (e.g., Valentine and Edmondson 2014).

How does it work? Typically, role-based coordination theories assert that work can be organized by assigning roles to individuals and holding them accountable for the responsibilities associated with their roles. Structure is used to coordinate work across roles and is determined by the relationships among roles within some boundary. Structure can be viewed as either a centralized hierarchical structure or a decentralized flat structure. Role-based coordination theories accomplish work by defining and assigning roles to individuals and ensuring that these roles are structured in a way that best supports the work needed to be done.

Role-Based Coordination Theory Key Elements

- Role definition
- Role assignment
- Role structure
- Role accountability

Potential Benefits for Macro-tasking: Role-based coordination does not rely on specific individuals to accomplish work but instead relies on roles. Reliance on roles promotes a plug-and-play structure that allows crowd workers to move in and out of the crowd with minimal disruption to work.

Potential Drawbacks for Macro-tasking: Role-based coordination requires someone to create the roles and their corresponding responsibilities. That being the case, it is not clear who would create new roles when needed. This becomes problematic when task requirements are emergent and change over time.

2.2.2.4 Relational Coordination Theory

What is it? Relational coordination theory asserts that a core facilitator of effective work is the quality of interactions among workers (Gittell 2002, 2011). According to Gittell (2002), the quality of interactions is based on effective communications and strong relationships. The underlying logic is that coordination involves both task interdependencies and the interactions among people involved in those tasks. Therefore, higher quality interactions among people involved in those tasks are likely to enhance coordination and lead to better performance (Gittell 2011). According to relational coordination theorists, high-quality relationships are especially important to achieving better performance when work is complex, interdependent, and time-constrained (Faraj and Xiao 2006; Gittell 2002, 2006, 2011). The importance of the relationships among employees has been supported by several observations in organizational settings (e.g., Adler et al. 2008).

How does it work? Relational coordination theory views coordination as “a mutually reinforcing process of interaction between communication and relationships carried

out for the purpose of task integration” (Gittell 2002, p. 301). Relational coordination theory describes relationship in terms of three dimensions: shared goals, shared knowledge, and mutual respect. The theory describes communication in four dimensions: frequency, timeliness, accuracy, and problem-solving focus (Gittell 2002, 2006). Relational coordination occurs when work is coordinated “through high-quality communication, supported by relationships of shared goals, shared knowledge, and mutual respect” (Gittell 2016, p. 11). This indicates that collectives who have more frequent, timely, accurate, and problem-solving-focused communication can be expected to coordinate more effectively and ultimately perform better by having shared goals, shared knowledge, and mutual respect.

Relational Coordination Theory Key Elements

- Relationships
 - Shared goals
 - Shared knowledge
 - Mutual respect
- Communication
 - Frequent
 - Timely
 - Accurate
 - Problem-solving focus

Potential Benefits for Macro-tasking: Coordination via high-quality relationships is very flexible and robust, allowing crowds to adapt to new or emergent task requirements. It relies less on formal planning and more on the possibility of informal planning done by the crowd itself.

Potential Drawbacks for Macro-tasking: It takes time to develop high-quality relationships among crowd workers. However, it is unclear whether current crowdsourcing platforms are designed to support the development of high-quality relationships among crowd workers.

2.2.2.5 Stigmergic Coordination Theory

What is it? Stigmergic coordination can be described as coordination that occurs through changes in a shared or collective work product (Rezgui and Crowston 2018). The concept of stigmergy is derived from entomologists’ observations of social insects. Insects such as ants and termites leave traces (e.g., pheromones) while performing work, and such traces stimulate other insects to take subsequent actions (Heylighen 2015; Khuong et al. 2016). Examples of stigmergic coordination on the part of insects include termites building and repairing nests, and ants finding the shortest route to food (Heylighen 2016; Khuong et al. 2016). The concept of stigmergy has influenced the design of collaborative action such as free open-source

software development (Bolici et al. 2009, 2016), multi-agent systems (e.g., Valckenaers et al. 2004) and collective robotics (e.g., Holland and Melhuish 1999). These areas have applied the stigmergic coordination approach to the need for coordinating in dynamic and emergent environments without direct communication between workers and agents.

How does it work? Members of a given collective not only perform work but also leave traces of their work. This requires ensuring that those traces are visible to other members. Those other members interpret those traces to determine what has already been done. Based on this, and their knowledge of what has to eventually be done, they determine the work that needs to be done next. Finally, as they are performing their own work they leave traces behind for other members. The stigmergic process of coordination occurs across many tasks done by many workers. As a result, stigmergic coordination can occur without direct and explicit interactions among members of a collective (Heylighen 2016). Stigmergic coordination seems to operate, in part, based on the development of shared work norms and practices normally associated with communities of practice (Lave 1991, 2009; Lave and Wenger 1991), derived somewhat from Suchman's (1987) work on situated action.

Stigmergic Coordination Theory Key Elements

- Create traces
- Interpret traces
- Determine future actions based on traces

Potential Benefits for Macro-tasking: Stigmergic coordination relies on distribution cognition, which allows the crowd to self-organize. There is low reliance on specific individuals to accomplish or plan the work. This provides a relative plug-and-play structure for crowd workers from the same community of practice (i.e., shared work norms). Stigmergic coordination employs informal planning that is flexible, robust, and adaptive to new or emergent task requirements.

Potential Drawbacks for Macro-tasking: Crowd workers must share a common set of work norms and practices. Therefore, the plug-and-play structure only applies to members of the same or similar work collectives. In fact, stigmergic coordination might be the worst coordination approach when workers do not share a common set of work norms and practices. As such, it limits the potential set of crowd workers available to recruit from.

2.2.2.6 Integrative Coordination Framework

What is it? The integrative framework was put forth by Okhuysen and Bechky (2009), in part to help identify coordination mechanisms. Based on their literature review on coordination they identified five types of coordination mechanisms (plans and rules, objects and representations, roles, routines, and proximity) and three conditions needed for coordination (accountability, predictability, and a common understanding). Generally, Okhuysen and Bechky's (2009) integrative framework asserts

that the five types of coordination mechanisms promote coordination through supporting one or more of the three conditions. Specifically, their framework identifies which coordination mechanisms support which conditions.

How does it work? The integrative framework promotes coordination by identifying the types of mechanisms needed. If one assumes that accountability, predictability, and a common understanding are needed, then one could ensure that at least one mechanism is chosen to support each of them. Likewise, if coordination was still a problem, more mechanisms could be employed to help buttress a particular condition. For example, if collectives were struggling with accountability, the integrative framework could help to identify a mechanism that could be employed to improve accountability.

Coordination Mechanisms

Plans and rules: As one of the fundamental elements of coordination, “plans and rules” refers to a set of elements that define relationships among tasks, workers, and other units of organizations. Among the functions of plans and rules is *defining responsibility for tasks*. Coordination by plans and rules enables people to decide what (subsequent) actions to take and what choices to make among the alternatives to complete tasks.

Objects and representations: The effective use of objects, representations, and technologies helps in coordinating work by providing information that is important to accomplish tasks (*direct information-sharing*). For example, boundary objects (e.g., data spreadsheets) are necessary to communicate problems to solve, ideas, and activities across teams. Also, a representative map or matrix of tasks and responsibilities (*scaffolding*) serves as a frame that reminds people of what tasks to do, the actors in charge of each task, the alignment of tasks among workers, and the progress of work (*acknowledging and aligning work*).

Roles: Roles can function as a coordination mechanism in two ways. While representing sets of responsibilities and activities of an actor who occupies the position, roles at once allow for redefining the responsibilities to adapt to the emergent status of work (*monitoring and updating*). This process of defining roles allows for *creating a common perspective*. Under common understandings about responsibilities, *substitution* can be easily done.

Routines: In more traditional organizational contexts, “routines” refers to “repeated patterns of behavior that are bound by rules and customs” (Feldman 2000, p. 611). In contrast, the current literature defines “routines” as ways to reflect “social meaning and social interaction ... embedded within them” (Okhuysen and Bechky 2009, p. 477).

Proximity: “Proximity” refers to coordination based on factors often associated with physical distance. These factors include *visibility* and *familiarity*. “Visibility” refers to the ability to see what others are doing, which is often associated with collocation but not necessarily a requirement of collocation. “Familiarity” refers to the ability to rely on prior relationships with others to facilitate the coordination of actions. Once

again, familiarity has often been associated with collocation but is not necessarily a requirement of collocation.

Conditions

Accountability: Accountability describes who is responsible for specific tasks and elements of those tasks. Making clear and visible who is in charge of which tasks promotes the awareness of each person's interdependence and responsibility, and the development of trust, which is in turn expected to contribute to coordinated actions in a collective. Accountability in the integrative framework includes the means that are created through informal and emergent interactions such as side conversations. Plans, rules, and objects can serve as the scaffolding that links tasks with people who are responsible for them. Roles, routines, and visibility also support continual monitoring, updating, and hand-offs among workers.

Predictability: Predictability explains workers' understanding of what subtasks constitute larger tasks in what sequence and what activities must be performed to accomplish each task. Predictability is essential for coordination because it highlights the anticipation of subsequent tasks and related actions of others and allows workers to adjust their work to others' work and perform their work accordingly. Plans and objects are the coordination mechanisms that create predictability by determining what tasks need to be completed. Familiarity and routines also enhance predictability by providing information on other workers' preferences with regard to the work.

Common understanding: Common understanding is a shared knowledge among workers about what the whole completed work is like, including goals and objectives and how it is accomplished. Plans and rules create a common understanding of the whole interdependent task and the process, facilitating better coordination. Routines and familiarity help workers become familiar with the ways the different parts of the work are put together to create the whole. In addition, objects and roles develop a common perspective through sharing and learning different activities to complete tasks.

Integrative Coordination Framework Key Elements

- Coordination mechanisms
 - Plans and rules
 - Objects and representations
 - Roles
 - Routines
 - Proximity
- Conditions
 - Accountability
 - Predictability
 - Common understanding

Potential Benefits for Macro-tasking: Because the integrative perspective entails both formal and emergent processes of coordination, the development of the coordination mechanisms and conditions promotes diverse coordination activities. This includes the explanation of a range of coordination procedures and tasks, from defining problems and tasks to completing and handing off tasks.

Potential Drawbacks for Macro-tasking: Establishing such mechanisms and conditions might require a specific set of personnel, which would be expected to take enough time to develop alternative formal and informal patterns of coordinated activities.

2.3 Recent Studies on Coordination in Macro-tasking Crowdsourcing

2.3.1 Search Methods

To review recent studies of coordination in macro-tasking, we first employed the academic search engine Google Scholar, entering the search keywords “microtask,” “coordination,” and “crowdsourcing.” We conducted the search in August 2018 and the results showed 60 articles. We read abstracts of the articles and evaluated whether to include the articles in the literature review based on the following inclusion criteria: (1) the article addressed issues about coordination for macro-tasking or (2) the article suggested and tested empirical ideas or designs of macro-task crowdsourcing. We excluded review papers, textbook-type books, patent applications, and articles published in non-English venues. Eight studies met all the criteria from the initial search. Additionally, we traced back some of the initial search results. This was because we found that some studies had been influencing the literature in macro-tasking coordination but had not shown up through our keyword search. For example, Kittur et al. (2011) and Kulkarni et al. (2012) were heavily cited as exemplar of investigating coordination problems of macro-tasking but didn’t appear in the initial search results. As a result, we identified a total of ten studies for the literature review.

2.3.2 Approaches Used to Coordinate Crowdsourcing Macro-tasks

We reviewed all the papers to identify which coordination theories and which of the five mechanisms were employed. To do this, we first grasped the main ideas and assumptions behind each coordination theory. We used these to make distinctions among them. Then we read and reviewed each study independently and discussed

which theory best represented each study's approach to coordination and whether it relied on one of the five mechanisms.

Most studies could be placed within the coordination theory approach (see Table 2.1). These studies typically focused on identifying and managing various dependencies among tasks, roles, and workers. To identify and manage dependencies these studies leveraged various techniques and tools. For example, to understand dependencies at the task level, Kittur et al. (2011) and Kulkarni et al. (2012) proposed systems that displayed plans for the work, including the sequence and the structure of work in units of subtasks. Also, to coordinate available competent workers, Haas et al. (2015) and Schmitz and Lykourantzou (2018) devised systems to model the structure of work by workers' level of skills and expertise. It appears that many HCI and CSCW researchers have addressed issues of coordination in macro-tasking, exploring the ideas best represented by coordination theory.

The second most used theory was role-based coordination. We found several studies that employed role-based coordination. These studies typically created a structure of roles and responsibilities for those roles and assigned qualified workers to each role to achieve goals. For example, Valentine et al. (2017) first built a hierarchical structure of roles based on tasks and activities using the role-based coordination theory. This study is in line with previous studies on scaffolding structures of roles in emergent coordination contexts. This includes an emergency unit of a university hospital (Valentine and Edmondson 2014) and emergent student team projects (Retelny et al. 2014; Valentine et al. 2017). We found no studies employing TMS or stigmergic coordination.

Regarding relational coordination, Salehi et al. (2017) study aligned with the relational coordination approach. The authors identified that familiarity among workers was an advantageous condition in performing tasks for distributed crowds. Specifically, when teaming workers up, they accounted for familiarity (e.g., history of collaborations with other members) in addition to availability. They also provided an instant communication channel and collaborative writing platform to support collaboration. The results indicated that the workers working with familiar teammates performed better, knowing well other team members' strengths and work processes. This study was not conducted in the same context as the face-to-face organization interaction that extant research in relational coordination has considered. However, by convening workers who were familiar with one another and leveraging their shared knowledge with the use of proper communication tools, the study successfully examined the effectiveness of relational communication.

In summary, it appears that scholars are overwhelmingly employing coordination theory to explore ways to handle macro-tasking in crowdsourcing. Role-based coordination is a distant second, followed by the relational coordination theory. None of the studies employed TMS or stigmergic approaches. Nonetheless, the literature base is quite nascent, with just two papers before 2015 (in 2011 and 2012) and more than half published in 2017 or 2018.

Table 2.1 Literature review based on coordination theories

	Transactive memory system	Coordination theory	Role-based coordination	Stigmergic coordination	Relational coordination
Kittur et al. (2011) (Crowdforge)		X			
Kulkarni et al. (2012) (Turkomatic)		X			
Haas et al. (2015) (Argonaut)		X	X		
Teevan et al. (2016) (Microwriter)		X			
Kim et al. (2017) (Mechanical novel)		X			
Retelny et al. (2017) (No workflow)		X			
Salehi et al. (2017) (Huddler)		X			X
Valentine et al. (2017) (Flash organization)			X		
Kaur et al. (2018) (Vocabulary)		X			
Schmitz and Lykourantzou (2018) (Task assignment and sequencing)		X			

2.3.3 *Coordination Mechanisms for Crowdsourcing Macro-tasks*

2.3.3.1 Evolving Plans and Rules

Plans and rules have been employed to help identify what tasks need to be completed and to assign crowd workers task responsibilities. Especially in the macro-tasking context, plans and rules for crowd workers should evolve to actively react to changes as work progresses. For example, Kulkarni et al. (2012) proposed Turkomatic, a real-time editable workflow that can be formed by crowds. Turkomatic was developed to allow workers to breakdown complex problems into smaller tasks. Kim et al. (2017) suggested a reflect-and-revise technique with which crowds could work on solving complex problems such as story-writing. Emphasizing the importance of higher level goals for complex and open-ended work, they utilized top-down goals for completing story-writing tasks. While the goals served to effectively accommodate outputs from different crowd workers, one distinct characteristic of this method was that goals were not pre-embedded in the writing system but were chosen among other workers from previous stages. Thus moving around the iterative steps of reflection and revision goals, workers came up with better ideas for given tasks.

2.3.3.2 Dynamic Objects and Representations

As a strategy of employing the objects and representations mechanism, workflows have been dominantly used in the crowdsourcing literature. Workflows serve as an object and representation that reflects the division and sequence of work. In macro-task coordination, because of the nature of macro-tasks—which are often non-decomposable, context-dependent, and contingent on progress and changes—designing workflows has been a challenging problem (Retelny et al. 2017).

Researchers have investigated workflows for macro-tasking that can be collaboratively developed and amenable to work progress. One example is Turkomatic, developed by Kulkarni et al. (2012). The system employs a list view and hierarchical graphs to show the structure of decomposed tasks by workers and the status of each task, whether waiting, in progress, or done. Another example is a sentence-level scaffolding structure that Kim et al. (2017) utilized to define subsequent goals and tasks to accomplish in Mechanical Novel. It helped workers not only generate suggestions for further edits on a draft but also identify goals and tasks at a given stage.

Objects have also been suggested to support workers in decomposing complex tasks. Kaur et al. (2018) introduced a “cognitive scaffold” for crowd workers to plan action items to accomplish complex and context-embedded tasks. Specifically, the researchers provided a vocabulary that comprised possible functions and sub-tasks based on the analysis of the crowd’s comments on possible writing goals. The researchers found it useful for workers to map out writing tasks.

2.3.3.3 Roles Loosely Held

We found several studies employing role-based coordination along with defining hierarchical role structures. Haas et al. (2015) built Argonaut, which automatized control of crowd workers' output and their quality. To review task output and quality effectively, the researchers defined positions of reviewers, reflecting different levels of their review expertise, and made a hierarchical structure of the positions. Using the hierarchy, the researchers identified a pool of trusted workers and assigned them to different positions. Valentine et al. (2017) proposed flash organizations that were flexibly assembled, role-based structures. The hierarchical structure loosely defined roles and responsibilities to help workers use their skills and competence to adjust to the progress of work. This approach allowed for the mobilization of different sets of crowd workers depending on their expertise and availability. In addition, to do more efficient substitution, Salehi et al. (2017) addressed the role mechanism by managing familiarity and availability. By creating a loosely bounded team that consisted of crowd workers who had a common understanding of their role and relationship to the project, the researchers could occupy roles with different workers who were available at a given point, and the researchers found that this approach supported complex-task completion.

2.3.3.4 Routines

We found one article that discussed the use of routines as a coordination mechanism. Salehi et al. (2017) noted that routines can be useful when uncertainty and complexity of a problem is low. As they noted, routines can help workers develop common knowledge about how to produce a desired outcome based on prescribed procedures. Salehi et al. (2017) discovered that worker familiarity, as routines would accomplish, could lead to better coordination by increasing workers' knowledge of how their teammates worked.

2.3.3.5 Proximity

Our review found one study that employed proximity as familiarity (see Salehi et al. 2017), but none employed proximity as visibility to coordinate macro-tasks. This might be because the studies we reviewed were motivated to tackle problems related to online crowdsourcing, where crowd workers are distributed and rarely have familiarity with one another.

Table 2.2 Literature review based on coordination mechanisms

	Plans and rules	Objects and representations	Roles	Routines	Proximity
Kittur et al. (2011) (Crowdforge)	X	X			
Kulkarni et al. (2012) (Turkomatic)		X			
Haas et al. (2015) (Argonaut)	X		X		
Teevan et al. (2016) (Microwriter)		X	X		
Kim et al. (2017) (Mechanical novel)	X	X			
Retelny et al. (2017) (No workflow)		X			
Salehi et al. (2017) (Huddler)			X	X	X
Valentine et al. (2017) (Flash organization)		X	X		
Kaur et al. (2018) (Vocabulary)	X	X			
Schmitz and Lykourantzou (2018) (Task assignment and sequencing)	X	X			

2.3.4 Summary

Overall, our review of coordination in the macro-tasking crowdsourcing literature revealed that much of the literature has focused on a small subset of coordination mechanisms. More specifically, we found that macro-tasking studies on coordination have largely focused on establishing plans and rules (80%) to describe a final goal and subtasks (see Table 2.2). This was followed by the studies on building objects and representations (50%). Role-based approaches were also used as a coordination mechanism for macro-tasks in a few studies (40%). Routines and proximity were discussed in one study.

2.4 Agenda for Future Research

Based on our brief literature review on coordination theories used in macro-tasking, the stigmergic and relational coordination theories have been studied the least, along with two integrative mechanisms: proximity (visibility and familiarity) and routines. Yet, we believe these theories and mechanisms offer the greatest potential for the crowdsourcing of macro-tasks. First, these theories and mechanisms rely on social processes of interaction along with adjustment to emergent states. They place much less emphasis on a priori definition of interdependencies among tasks or even roles among crowd workers. Approaches that focus on defining work upfront are likely to always rely heavily on requestors. To the contrary, both stigmergic and relational coordination along with proximity (visibility and familiarity) and routines rely more on facilitating the establishment of more informal coordination, which allows for more spontaneous coordination of work. We believe these informal coordination approaches are likely to be more effective ways of coordinating crowdsourcing as it becomes increasingly oriented to macro- rather than micro-tasks. In addition, many of the concepts of TMS are embodied in relational coordination's "shared knowledge" concept.

To advance our understanding in the areas of stigmergic and relational coordination, we present and discuss several important research questions. In addition, we present design propositions related to stigmergic and relational coordination. Design propositions are general statements regarding the relationship between a design element and other concepts. In this chapter, design propositions are general statements regarding the relationship between the design of a system and coordination approaches.

2.4.1 Stigmergic Coordination

Stigmergic coordination refers to coordination based on traces, without explicit communication among workers (Heylighen 2016; Rezgui and Crowston 2018). Because stigmergic coordination doesn't necessarily require communication among workers and is done instead by interaction between workers and environments, including traces left by other workers, it could be beneficial in coordinating macro-tasks. For example, the stigmergic coordination process doesn't involve setting up plans and controls. This would help crowd workers readily get involved in work and adjust their behaviors to the status and progress of work. Thus, we suggest research questions that could advance macro-task coordination by employing stigmergic approaches.

Research Question 1: How can we support the traces of prior work in the crowdsourcing of macro-tasks?

First, as discussed, traces in stigmergic coordination serve as mediating objects that enable the bridging of the actions of prior workers with those of subsequent workers.

Traces help inform workers of both the progress of work and the remaining work. Thus, developing systems that support leaving traces effectively could be one way to support stigmergic coordination for crowdsourcing macro-tasks. For example, crowdsourcing systems could be designed to provide features that help workers leave comments or remarks next to their work. These systems could be designed to include features that track the progress of work and make it salient.

Research Question 2: How can we promote the shared interpretations of traces in the crowdsourcing of macro-tasks?

Workers who engage in stigmergic coordination use traces to implicitly determine what has been done and what to do next. This implicit coordination can occur because the workers belong to a community that has a shared context. This shared context helps to establish common work norms and routines among members of a given community. This is what allows workers to employ traces as a mechanism to engage in implicit coordination. Next, we discuss three approaches to leveraging stigmergic coordination in macro-tasking crowdsourcing.

One approach is to recruit crowd workers who already have a shared context, norms, and routines. This could be done by recruiting groups of workers from existing online communities and peer platforms like GitHub. For example, a group of workers from GitHub could be recruited to work on a macro-tasking project. These workers would already have a shared context, norms, and routines. To leverage their existing shared context, norms, and routines obtained using the GitHub platform, the crowdsourcing platform should be set up similarly to the GitHub platform. Together the workers from the GitHub community and the new crowdsourcing platform that supports the workers' shared context, norms, and routines should allow crowd workers to engage in stigmergic coordination to tackle macro-tasks.

Another approach is to create an online community from which to recruit crowd workers. This approach offers two advantages. One, it would allow crowd workers to develop a shared context, norms, and routines. Over time, these crowd workers would be able to engage in stigmergic coordination in the same way as crowd workers who are members of current online communities. Two, this approach would allow for the creation of an online community that focuses on a subject or theme that might not exist. For example, imagine if macro-tasks required workers who were familiar with a specific programming language like the common business-oriented language (COBOL). Many mainframes still rely on programs written in this language, although it is not widely taught. Creating an online community of COBOL programmers would support recruitment for macro-tasks requiring COBOL.

Finally, the third approach is to require crowd workers who want to participate in macro-tasking to have experience working in a specific online community. Potential workers would be directed to participate in a specific online community before they could be eligible to be selected for macro-tasking. This would allow crowd workers the opportunity to learn basic knowledge and rules from an existing online community. Over time they would develop the shared context, norms, and routines needed to be selected for macro-tasks.

Table 2.3 Design propositions for stigmergic coordination

Stigmergic coordination design propositions
Design proposition 1: Crowdsourcing systems that support stigmergic coordination will help crowd workers effectively accomplish macro-tasks
Design proposition 1a: To promote stigmergic coordination, crowdsourcing systems must facilitate the leaving and making visible the traces of prior work
Design proposition 1b: To promote stigmergic coordination, crowdsourcing systems must facilitate a shared interpretation of the traces of prior work
Design proposition 1c: To promote stigmergic coordination, crowdsourcing systems must support the leveraging of shared work norms and practices

Table 2.3 presents a summary of the three design propositions related to stigmergic coordination. Design propositions were derived from the research questions 1 and 2.

2.4.2 Relational Coordination

Relational coordination theory describes relationship in terms of three dimensions: shared goals, shared knowledge, and mutual respect; and communication in four dimensions: frequency, timeliness, accuracy, and problem-solving focus. These dimensions are both representative of and impacted by the quality of social relationship within a given collective. The benefits of relational coordination are that it allows workers to coordinate complex work in dynamic environments. This is accomplished by allowing individuals to coordinate their efforts by working through problems cooperatively. Relational coordination can be viewed as a set of mechanisms that provide a canvas for a collective set of painters. As long as collectives maintain quality relationships, they can leverage elements of their relationships to effectively coordinate work. In fact, it is this reliance on the quality of relationships that clearly differentiates relational coordination from stigmergic coordination.

Next, we suggest research questions that could advance our understanding of crowdsourcing macro-tasks through relational coordination.

Research Question 3a: How can shared knowledge be promoted in the crowdsourcing of macro-tasks?

According to relational coordination, shared knowledge helps workers to become aware of their interdependencies with coworkers and of one another's potential contribution to work. This awareness helps to facilitate effective and accurate communication. There are two big challenges with achieving a sufficient level of shared knowledge in crowdsourcing. One, workers engaged in crowdsourcing are often ad hoc and have little prior experience working together. Therefore, they initially have little or no shared knowledge as a group. Two, depending on the amount of time required to complete the task, crowd workers often do not have enough time to

develop shared knowledge. Both challenges greatly undermine the ability of crowd workers to rely on shared knowledge as a coordination mechanism.

There are several potential ways to design crowdsourcing systems to promote shared knowledge. First, systems could help crowd workers identify who knows what. This could be done by having a system that publicly displays each worker's profile. This profile could include the worker's educational and work experience. The workers should give consent before profiles are displayed, and more or less information might be displayed based on who is viewing the profile. For example, members of the macro-task team might have access to more information about each worker than members of the public. Second, systems should be designed to help make as much as possible of the individual crowd worker's knowledge explicitly available to all others. This could be done by promoting the sharing, using, and ultimate integration of knowledge across the team (Robert et al. 2008, 2018). Crowdsourcing systems would need to be designed to not only provide both asynchronous and synchronous communication capabilities but several other important features. For example, these systems should make it easy to search the repository of communications, including multichannel communications and use of visual aids such as sketches, snapshots, whiteboards, links, documents, and templates (Alavi and Tiwana 2002). These features should also provide real-time editing and commenting so that workers could explain their actions to others as well as inquire about why actions were taken.

Research Question 3b: How can shared goals be leveraged in the crowdsourcing of macro-tasks?

Shared goals are another important coordination mechanism in relational coordination that can be problematic in crowdsourcing macro-tasks. Shared goals motivate workers to engage in high-quality communication with others. This guides workers to focus more on problem-solving-related communication than emotional and non-productive communication. On one hand, it should be easy to promote shared goals in the crowdsourcing of macro-tasks. The crowd workers have been assembled to accomplish a specific macro-task. This macro-task is essentially the shared goal. On the other hand, it can be difficult for crowd workers to maintain a shared view on the progress or lack of progress of those shared goals. This can be even more problematic in macro-task work environments, which can be more dynamic than static micro-tasking work environments.

To promote a shared view of goals in the crowdsourcing of macro-tasks, we turn to boundary objects. According to Okhuysen and Bechky (2009), boundary objects are a type of object and representation coordination mechanism. As stated, boundary objects help to communicate problems, ideas, and activities across teams. The biggest benefit of boundary objects is that they allow an individual's specific understanding of a given situation to be framed within the larger context of the collective's situation (Bechky 2003). Therefore, boundary objects can be used to communicate the status of the collective's situation to all members of the collective, without the need for workers to fully understand each member's specific situation. In the case of crowdsourcing macro-tasks, boundary objects could promote a shared view of goals by allowing crowd workers to accomplish individual objectives within

the framework of the collective's goals. However, it is not clear which boundary objects should be employed. One option would be to focus on promoting situation awareness.

The promotion of situation awareness offers a viable approach to understanding how to design boundary objects to promote a shared view of goals in macro-tasks. Endsley (1995) formally defined situation awareness as “the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future” (p. 36). A more informal definition is an ability to perceive and comprehend information, which allows for the prediction of future courses of action in a dynamic environment. In the case of crowdsourcing, we define crowdsourcing situation awareness as the ability of crowd workers to perceive and comprehend the status of their crowd's work and to forecast the needed future courses of action to complete the crowd's work. Situation awareness is similar to the use of traces in stigmergic coordination, with several differences. The use of traces in stigmergic coordination is the result of a shared context, norms, and routines obtained in large part by one's socialization into a community. However, situation awareness can be obtained without the need for this socialization process, and although it can help to promote a shared context, it may or may not lead to shared norms and routines. In addition, stigmergic coordination occurs through implicit coordination, whereas situation awareness occurs primarily through explicit coordination among members of the collective.

How can systems be designed to support situation awareness in the crowdsourcing of macro-tasks?

To accomplish this, scholars should turn to the study of visualization. Visualization is science that focuses on understanding how to best display information to humans. A full review of this research area is beyond the scope of this book chapter, but visualization techniques have been used to reduce cognitive load (Anderson et al. 2011). It is likely that current research on visualization can be leveraged and that new research will also be needed. Questions about how best to support situation awareness specifically for crowdsourcing macro-tasks would need to be addressed. A program of research in this area might attempt to define key attributes of the type of macro-task and crowd workers, and stage of work, and how these factors influence the ways information should be displayed.

Research Question 3d: How can mutual respect be promoted in the crowdsourcing of macro-tasks?

In relational coordination, mutual respect increases the level of receptiveness to communication with others, leading to increased opportunity for improving shared knowledge and solving problems effectively. On one hand, the challenges to achieving mutual respect are the same as those to achieving shared knowledge in crowdsourcing macro-tasks. These include the often ad hoc nature of crowdsourcing, which involves assembling crowd workers with little experience working together and a short duration of time required to complete the task. Some challenges are also different; for example, crowd workers could also develop a mutual disrespect for one

another. Each of these challenges could greatly undermine the ability of crowd workers to rely on mutual respect as a coordination mechanism.

To combat these challenges, there are several potential ways to design crowdsourcing systems to promote mutual respect. First, systems could promote mutual respect through trust. This could be done by designing systems that display recommendations from others who have worked with the crowd worker. This system could share positive narratives about the crowd worker's behavior. Such a system could include a peer evaluation that rates crowd workers on their respect for others. Second, systems could be designed to monitor the level of mutual respect among crowd workers. For example, Munson et al. (2014) developed a system that monitored the email communications among teams to determine their degree of trust and respect through linguistic mimicry. Questions around how such systems could measure mutual respect or what data should be used to measure it would need to be further investigated. For example, it is not clear how such measures might be drawn from prior studies or whether new measures better suited to a macro-tasking context need to be identified. Systems like these could be designed to diagnose the level of mutual respect among crowd workers to determine whether interventions are needed.

Finally, interventions should be designed to help promote mutual respect when needed. Although research is needed to understand the types of interventions necessary, we recommend several potential avenues. The research on conflict and conflict resolution offers a rich set of literature to draw from. For example, this research has identified several types of conflict: relationship, process, and task (Jehn 1997). Relationship conflict is related to personal disagreements among team members, whereas task- and process-focused conflicts are related to work but are not personal disagreements. Research has shown that relationship conflict is always detrimental to performance, whereas task and process conflicts can be beneficial to team performance (Windeler et al. 2015). Systems should be designed to determine which type of conflict is occurring. The literature on conflict resolution has identified several approaches to resolving conflict in groups. These include avoidance, accommodation, competition, collaboration, and compromise (Kankanhalli et al. 2006; Montoya-Weiss et al. 2001; Paul et al. 2004). Although a full review and discussion of each of these are beyond the scope of this chapter, what is clear is that each approach has pros and cons and would likely require different system interventions. A program of research could explore both the effectiveness of each approach in the context of crowdsourcing macro-tasks and how to best design systems to support each approach.

Research Question 3e: What is the most effective way to promote communication in the crowdsourcing of macro-tasks?

Relational coordination defines communication in four dimensions: frequency, timeliness, accuracy, and problem-solving focus (Gittell 2002, 2006). The easiest and first step toward supporting frequent, timely, accurate, and problem-solving-focused communication is to design crowdsourcing systems that allow effective communications. Features of such systems have been identified in the form of both asynchronous and synchronous communications as well as multichannel communications. However, systems could be designed to go beyond this and take a more active role in several

Table 2.4 Design propositions for relational coordination

Relational coordination design propositions
Design proposition 2: Crowdsourcing systems that support relational coordination will help crowd workers effectively accomplish macro-tasks
Design proposition 2a: To promote relational coordination, crowdsourcing systems must facilitate the creating and sharing of collective knowledge
Design proposition 2b: To promote relational coordination, crowdsourcing systems must facilitate the creating and sharing of common goals
Design proposition 2c: To promote relational coordination, crowdsourcing systems must support the development of mutual respect
Design proposition 2d: To promote relational coordination, crowdsourcing systems must facilitate effective communication

meaningful ways. Systems could be designed to prompt communications. A research agenda could be built on the investigation of the effectiveness of types of prompts. For example, days before a work deadline the system could send an email to everyone inquiring about the status of the group's work. This might encourage crowd workers to engage in task-focused communications about the upcoming deadline. Nudges could also be used to alert crowd workers when the status of their group's work has changed or when crowd workers have left questions for others to answer. Crowdsourcing systems could be set up to require timely status updates that rely on the input of every crowd worker and go out to every crowd worker. A research agenda could also be built on understanding the effectiveness of the content of such messages. For example, research has shown that the framing of messages impacts how people choose to respond or not respond to them (Jung and Mellers 2016). Research should be directed at understanding the best content to promote communication frequency, timeliness, accuracy, and problem-solving focus among crowd workers.

Table 2.4 presents a summary of the four design propositions related to relational coordination. Design propositions were derived from research questions 3a, 3b, 3c, 3d and 3e.

2.4.3 *Limitations*

In this chapter, we acknowledge that theories of coordination have shared or overlapping concepts. Nonetheless, for the most part, we treated them as separate and distinct when discussing their pros and cons. Our separation of each theory of coordination might at times have been more artificial and arbitrary. Scholars studying issues related to crowdsourcing coordination should consider hybrid approaches that combine various elements of each theory. For example, stigmergic coordination could be augmented with role-based coordination. This could be accomplished by bringing in outsiders unfamiliar with the work norms and practices and defining a specific

role for them in the work structure. By defining their role, work disruption resulting from their lack of familiarity with traces should be kept at a minimum. We also acknowledge that each theory has its own rich and insightful literature that goes beyond the scope of this one chapter. This chapter provides a brief introduction of each theory. Where brevity and conciseness end and confusion and incompleteness begin is often debatable. That being the case, the goal of this chapter was to draw attention to the issues related to coordinating macro-tasking in crowdsourcing environments. Our recommendations are but suggestions and readers are advised to dig deeper into these issues themselves. Finally, we provide design propositions that link theory to design elements. Our propositions, like all propositions, are general statements. Ultimately, hypotheses should be derived from our design propositions before they can be empirically tested. This is a challenge we hope future scholars choose to undertake.

2.5 Conclusions

Crowdsourcing macro-tasking places more emphasis on coordinating complex, interdependent, and less decomposable tasks. This chapter reviewed and recommended several theories of coordination to address issues related to coordinating macro-tasks. It presented a research agenda and design propositions for each recommended theory of coordination. The research agendas and design propositions are far from complete, and more work is needed with regard to both theoretical development and empirical verification. Nonetheless, we hope this chapter is the first step in advancing our understanding of crowdsourcing coordination used for macro-tasks.

Acknowledgements This book chapter was supported in part by the National Science Foundation [grant CHS-1617820].

References

- Adler, P. S., Kwon, S. W., & Heckscher, C. (2008). Perspective—professional work: The emergence of collaborative community. *Organization Science*, *19*(2), 359–376.
- Alavi, M., & Tiwana, A. (2002). Knowledge integration in virtual teams: The potential role of KMS. *Journal of the American Society for Information Science and Technology*, *53*(12), 1029–1037.
- Anderson, E. W., Potter, K. C., Matzen, L. E., Shepherd, J. F., Preston, G. A., & Silva, C. T. (2011). A user study of visualization effectiveness using EEG and cognitive load. In *Computer graphics forum* (Vol. 30, No. 3, pp. 791–800). Oxford, UK: Blackwell Publishing Ltd.
- Andres, H. P., & Zmud, R. W. (2002). A contingency approach to software project coordination. *Journal of Management Information Systems*, *18*(3), 41–70.
- Austin, J. R. (2003). Transactive memory in organizational groups: The effects of content, consensus, specialization, and accuracy on group performance. *Journal of Applied Psychology*, *88*(5), 866–878.
- Bechky, B. A. (2003). Sharing meaning across occupational communities: The transformation of understanding on a production floor. *Organization Science*, *14*(3), 312–330.

- Bechky, B. A. (2006). Gaffers, gofers, and grips: Role-based coordination in temporary organizations. *Organization Science*, 17(1), 3–21.
- Bolici, F., Howison, J., & Crowston, K. (2009). *Coordination without discussion? Socio-technical congruence and stigmery in free and open source software projects*. Paper presented at the International Conference on Software Engineering, Vancouver, BC, Canada. Retrieved from <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.193.7473&rep=rep1&type=pdf>.
- Bolici, F., Howison, J., & Crowston, K. (2016). Stigmery coordination in FLOSS development teams: Integrating explicit and implicit mechanisms. *Cognitive Systems Research*, 38, 14–22.
- Brandon, D. P., & Hollingshead, A. B. (2004). Transactive memory systems in organizations: Matching tasks, expertise, and people. *Organization Science*, 15(6), 633–644.
- Choi, S. Y., Lee, H., & Yoo, Y. (2010). The impact of information technology and transactive memory systems on knowledge sharing, application, and team performance: A field study. *MIS Quarterly*, 34(4), 855–870.
- Crowston, K. (1994). A taxonomy of organisational dependencies and coordination mechanisms. MIT Center for Coordination Science Working Paper. Massachusetts Institute of Technology, August 1994.
- Crowston, K., Howison, J., & Rubleske, J. (2006). Coordination theory: A ten year retrospective. In P. Zhang & D. Galletta (Eds.), *Human-computer interaction in management information systems—foundations* (pp. 120–138). Armonk, NY: M. E. Sharpe Inc.
- Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors*, 37(1), 32–64.
- Faraj, S., & Xiao, Y. (2006). Coordination in fast-response organizations. *Management Science*, 52(8), 1155–1169.
- Feldman, M. S. (2000). Organizational routines as a source of continuous change. *Organization Science*, 11(6), 611–629.
- Gittell, J. H. (2002). Coordinating mechanisms in care provider groups: Relational coordination as a mediator and input uncertainty as a moderator of performance effects. *Management Science*, 48(11), 1408–1426.
- Gittell, J. H. (2006). Relational coordination: Coordinating work through relationships of shared goals, shared knowledge and mutual respect. In O. Kyriakidou & M. F. Özbilgin (Eds.), *Relational perspectives in organizational studies: A research companion* (pp. 74–94). Cheltenham, UK: Edward Elgar Publishers.
- Gittell, J. H. (2011). New directions for relational coordination theory. In K. S. Cameron & G. M. Spreitzer (Eds.), *The Oxford handbook of positive organizational scholarship* (pp. 400–411). New York, NY: Oxford University Press.
- Gittell, J. H. (2016). *Transforming relationships for high performance: The power of relational coordination*. Palo Alto, CA: Stanford University Press.
- Haas, D., Ansel, J., Gu, L., & Marcus, A. (2015). Argonaut: Macrotask crowdsourcing for complex data processing. *Proceedings of the VLDB Endowment*, 8(12), 1642–1653.
- Heylighen, F. (2015). Stigmery as a universal coordination mechanism: Components, varieties and applications. In T. Lewis & L. Marsh (Eds.), *Human stigmery: Theoretical developments and new applications*. New York, NY: Springer.
- Heylighen, F. (2016). Stigmery as a universal coordination mechanism I: Definition and components. *Cognitive Systems Research*, 38, 4–13.
- Holland, O., & Melhuish, C. (1999). Stigmery, self-organization, and sorting in collective robotics. *Artificial Life*, 5(2), 173–202.
- Howe, J. (2006). The rise of crowdsourcing. *Wired*, 14(6), 1–4.
- Jehn, K. A. (1997). A quantitative analysis of conflict types and dimensions in organizational groups. *Administrative Science Quarterly*, 42(3), 530–557.
- Jung, J. Y., & Mellers, B. A. (2016). American attitudes toward nudges. *Judgment & Decision Making*, 11(1), 62–74.
- Kankanhalli, A., Tan, B. C., & Wei, K. K. (2006). Conflict and performance in global virtual teams. *Journal of Management Information Systems*, 23(3), 237–274.

- Kaur, H., Williams, A. C., Thompson, A. L., Lasecki, W. S., Iqbal, S. T., & Teevan, J. (2018). Creating better action plans for writing tasks via vocabulary-based planning. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW), 86.
- Khuong, A., Gautrais, J., Perna, A., Sbaï, C., Combe, M., Kuntz, P., et al. (2016). Stigmergic construction and topochemical information shape ant nest architecture. *Proceedings of the National Academy of Sciences*, 113(5), 1303–1308.
- Kim, J., Sterman, S., Cohen, A. A. B., & Bernstein, M. S. (2017). Mechanical novel: Crowdsourcing complex work through reflection and revision. In *Proceedings of the 2017 ACM Conference on Computer-supported Cooperative Work and Social Computing* (pp. 233–245). New York, NY: ACM.
- Kittur, A., Smus, B., Khamkar, S., & Kraut, R. E. (2011). Crowdforge: Crowdsourcing complex work. In *Proceedings of the 24th annual ACM symposium on User Interface Software and Technology* (pp. 43–52). New York, NY: ACM.
- Kulkarni, A., Can, M., & Hartmann, B. (2012). Collaboratively crowdsourcing workflows with Turkomatic. In *Proceedings of the ACM 2012 Conference on Computer-supported Cooperative Work* (pp. 1003–1012). New York, NY: ACM.
- Lave, J. (1991). Situating learning in communities of practice. *Perspectives on Socially Shared Cognition*, 2, 63–82.
- Lave, J. (2009). The practice of learning. In K. Illeris (Ed.), *Contemporary learning theories* (pp. 200–208). London, UK: Routledge.
- Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. New York, NY: Cambridge University Press.
- Lewis, K. (2003). Measuring transactive memory systems in the field: Scale development and validation. *Journal of Applied Psychology*, 88, 587–604.
- Malone, T. W., & Crowston, K. (1994). The interdisciplinary study of coordination. *ACM Computing Surveys (CSUR)*, 26(1), 87–119.
- Marques-Quinteiro, P., Curral, L., Passos, A. M., & Lewis, K. (2013). And now what do we do? The role of transactive memory systems and task coordination in action teams. *Group Dynamics: Theory, Research, and Practice*, 17(3), 194–206.
- Montoya-Weiss, M. M., Massey, A. P., & Song, M. (2001). Getting it together: Temporal coordination and conflict management in global virtual teams. *Academy of Management Journal*, 44(6), 1251–1262.
- Moreland, R. L. (1999). Transactive memory: Learning who knows what in work groups and organizations. In L. L. Thompson, J. M. Levine, & D. M. Messick (Eds.), *Shared cognition in organizations: The management of knowledge* (pp. 3–31). Mahwah, NJ: Erlbaum.
- Munson, S. A., Kervin, K., & Robert Jr., L. P. (2014). Monitoring email to indicate project team performance and mutual attraction. In *Proceedings of the 17th ACM Conference on Computer-supported Cooperative Work & Social Computing* (pp. 542–549). New York, NY: ACM.
- Okhuysen, G. A., & Bechky, B. A. (2009). Coordination in organizations: An integrative perspective. In J. P. Walsh & A. P. Brief (Eds.), *Academy of management annals* (Vol. 3, pp. 463–502). Essex, UK: Routledge.
- Paul, S., Seetharaman, P., Samarah, I., & Mykytyn, P. P. (2004). Impact of heterogeneity and collaborative conflict management style on the performance of synchronous global virtual teams. *Information & Management*, 41(3), 303–321.
- Ren, Y., & Argote, L. (2011). Transactive memory systems 1985–2010: An integrative framework of key dimensions, antecedents, and consequences. *Academy of Management Annals*, 5(1), 189–229.
- Retelny, D., Bernstein, M. S., & Valentine, M. A. (2017). No workflow can ever be enough: How crowdsourcing workflows constrain complex work. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW), 89.
- Retelny, D., Robaszekiewicz, S., To, A., Lasecki, W. S., Patel, J., Rahmati, N.,... Bernstein, M. S. (2014). Expert crowdsourcing with flash teams. In *Proceedings of the 27th annual ACM Symposium on User Interface Software and Technology* (pp. 75–85). New York, NY: ACM.

- Rezgui, A., & Crowston, K. (2018). Stigmergic coordination in Wikipedia. In *Proceedings of the 14th International Symposium on Open Collaboration* (pp. 1–12). Paris, France: ACM Press.
- Robert, L. P. (2016). Far but near or near but far?: The effects of perceived distance on the relationship between geographic dispersion and perceived diversity. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 2461–2473). New York, NY: ACM.
- Robert, L. P., Dennis, A. R., & Ahuja, M. (2008). Social capital and knowledge integration in digitally enabled teams. *Information Systems Research*, 19(3), 314–334. <http://pubsonline.informs.org/doi/abs/10.1287/isre.1080.0177>.
- Robert, L. P., Dennis, A. R., & Ahuja, M. (2018). Differences are different: Examining the effects of communication media on the impacts of racial and gender diversity in decision-making teams. *Information Systems Research*, 29(3), 525–545. <https://doi.org/10.1287/isre.2018.0773>.
- Robert, L. P., & Romero, D. M. (2015). Crowd size, diversity and performance. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (pp. 1379–1382). New York, NY: ACM.
- Robert, L. P., Jr., & Romero, D. M. (2017). The influence of diversity and experience on the effects of crowd size. *Journal of the Association for Information Science and Technology*, 68(2), 321–332.
- Salehi, N., McCabe, M., Valentine, M., & Bernstein, M. S. (2017). Huddler: Convening stable and familiar crowd teams despite unpredictable availability. In *Proceedings of the 20th ACM Conference on Computer-supported Cooperative Work & Social Computing (CSCW'17)*. New York, NY: ACM.
- Schmitz, H., & Lykourantzou, I. (2018). Online sequencing of non-decomposable macrotasks in expert crowdsourcing. *ACM Transactions on Social Computing*, 1(1), 1–34.
- Strode, D. E., Huff, S. L., Hope, B., & Link, S. (2012). Coordination in co-located agile software development projects. *Journal of Systems and Software*, 85(6), 1222–1238.
- Suchman, L. (1987). *Plans and situated actions: The problem of human-machine communication*. Cambridge, MA: Cambridge University Press.
- Teevan, J., Iqbal, S. T., & Von Veh, C. (2016). Supporting collaborative writing with microtasks. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 2657–2668). New York, NY: ACM.
- Valckenaers, P., Kollingbaum, M., & Van Brussel, H. (2004). Multi-agent coordination and control using stigmergy. *Computers in Industry*, 53(1), 75–96.
- Valentine, M. A., & Edmondson, A. C. (2014). Team scaffolds: How mesolevel structures enable role-based coordination in temporary groups. *Organization Science*, 26(2), 405–422.
- Valentine, M. A., Retelny, D., To, A., Rahmati, N., Doshi, T., & Bernstein, M. S. (2017, May). Flash organizations: Crowdsourcing complex work by structuring crowds as organizations. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (pp. 3523–3537). New York, NY: ACM.
- Wegner, D. M. (1987). Transactive memory: A contemporary analysis of the group mind. In B. Mullen & G. R. Geothals (Eds.), *Theories of group behavior* (pp. 185–208). New York, NY: Springer.
- Windeler, J., Maruping, L., Robert, L. P., & Riemenschneider, C. (2015). E-identity, conflict and shared understanding in distributed teams. *Journal of the Association for Information Systems*, 16(7), 608–645.

Chapter 3

Crowdsourcing Controls: A Review and Research Agenda for Crowdsourcing Controls Used for Macro-tasks



Lionel P. Robert Jr.

Abstract Crowdsourcing—the employment of ad hoc online labor to perform various tasks—has become a popular outsourcing vehicle. Our current approach to crowdsourcing—focusing on micro-tasks—fails to leverage the potential of crowds to tackle more complex problems. To leverage crowds to tackle more complex macro-tasks requires a better comprehension of crowdsourcing controls. Crowdsourcing controls are mechanisms used to align crowd workers’ actions with predefined standards to achieve a set of goals and objectives. Unfortunately, we know very little about the topic of crowdsourcing controls directed at accomplishing complex macro-tasks. To address issues associated with crowdsourcing controls for macro-tasks, this chapter has several objectives. First, it presents and discusses the literature on control theory. Second, this chapter presents a scoping literature review of crowdsourcing controls. Finally, the chapter identifies gaps and puts forth a research agenda to address these shortcomings. The research agenda focuses on understanding how to employ the controls needed to perform macro-tasking in crowds and the implications for crowdsourcing system designers.

3.1 Introduction

Crowdsourcing—the employment of ad hoc online labor to perform various tasks—has become a popular outsourcing vehicle. Digital platforms like Mechanical Turk (<http://www.mturk.com>), CrowdFlower (<http://www.crowdflower.com>), MobileWorks (<http://www.mobileworks.com>), and Crowdcrafting (<http://crowdcrafting.org>) are in part responsible for the emergence and popularity of crowdsourcing. These popular platforms have been dominated by micro-tasks—standalone decomposed tasks (Schmitz and Lykourentzou 2018). This arrangement—micro-tasking through digital platforms—has been successful at providing organizations with access to affordable labor available 24 h a day (Ye et al. 2017).

L. P. Robert Jr. (✉)
University of Michigan School of Information, Ann Arbor, USA
e-mail: lprobert@umich.edu

Our current approach to crowdsourcing—focusing on micro-tasks—fails, however, to leverage the potential of crowds to tackle more complex problems. Addressing complex problems requires collaboration among individuals who hold multiple perspectives and diverse expertise. Crowdsourcing affords the opportunity to assemble individuals with a diversity of knowledge and skills that is not often available to a single individual or organization. However, employing this collective knowledge to tackle complex problems requires the shift from standalone micro-tasking to more collaborative macro-tasking. Macro-tasks are complex crowd work that is sometimes but not always decomposable to micro-tasks and requires collaboration among crowd workers to accomplish (Schmitz and Lykourantzou 2018).

Crowdsourcing controls are mechanisms used to align crowd workers' actions with predefined standards to achieve a set of goals and objectives. These goals and objectives are often set by the requestor, organization, or platform but can be set by the crowd itself. Crowdsourcing controls can be classified as those that influence the inputs, behaviors, and outputs of crowds and their workers. In the crowdsourcing literature, issues of control are usually addressed indirectly through individual financial incentives (Ye et al. 2017). Financial incentives used in crowdsourcing are designed to influence the effort and attention of crowd workers. This makes sense when crowds are performing individual standalone micro-tasks. This makes less sense for macro-tasks, which require group cooperation. Unfortunately, we know very little about the topic of crowdsourcing controls directed at groups (Daniel et al. 2018).

To address issues associated with crowdsourcing controls for macro-tasks, this chapter has several objectives. First, it presents and discusses the literature on control theory. This includes behavior–output control systems developed by Ouchi and the integrative model developed by Cardinal. These frameworks represent the most widely used control theories in the organizational behavior literature (Cardinal et al. 2017). Second, this chapter presents a scoping literature review that surveys the conceptualization and operationalization of crowdsourcing controls in the HCI/CSCW, information systems and organizational behavior literature. In doing so, this chapter highlights current approaches to controls used in crowdsourcing with an emphasis on what is needed to support macro-tasking. Finally, the chapter identifies gaps and puts forth a research agenda to address these shortcomings. The research agenda focuses on understanding how to employ the controls needed to perform macro-tasking in crowds and the implications for crowdsourcing system designers.

3.2 Background

3.2.1 *Micro-tasking Versus Macro-tasking Controls*

The first question one might ask is: *Why not employ controls used in micro-tasking to accomplish macro-tasking?* In other words, what makes macro-tasking so different that we need to rethink our approach to controls in crowdsourcing? Micro-

tasks are different from macro-tasks in the following ways. First, micro-tasks are already decomposed. Decomposition allows for crowd work to be transparent and predictable. Both transparency and predictability reduce the complexity associated with controls. Second, micro-tasks are standalone independent tasks that require little to no cooperation among crowd members. This narrows the problem of control to the actions of a single individual rather than a group. Third, micro-tasks are homogeneous with similar goals—multiple crowd workers are often performing the same task or set of tasks with the same or similar goal. This decreases the possibility of crowd members having conflicting goals and allows the same control to be used across the crowd.

Macro-tasks, on the other hand, are not decomposed, and in some cases cannot be decomposed. Therefore, crowd work for macro-tasking is often not very straightforward or predictable. This requires crowd workers to negotiate what needs to be done, and in some cases, this happens in real time. This introduces the problem of determining not only which controls to employ but also who should employ them. Second, macro-tasks are not standalone independent tasks but instead interdependent tasks requiring cooperation and coordination among crowd members. As such, the problem of controls requires understanding how to control the actions of a group—not just individuals. Third, macro-tasking requires crowds to undertake a diverse set of tasks, each with its own goals and objectives. Therefore, workers in the same crowd can have different goals associated with their part of the macro-task. This makes it much harder to align goals using a single control. As such, one control might be effective for one component of a macro-task but not another. Issues related to the use of multiple types of group controls in crowdsourcing have largely been ignored.

3.2.2 Control Theory in the Organizational Behavior/Science Literature

Control is viewed as one of the four primary functions of management (Carpenter et al. 2010). This is often embodied in the planning, organizing, leading, and controlling (PLOC) framework used in most basic management books. Controls are goal-oriented in that they direct employees' actions to a specific goal, and controls are multifaceted in that there is a diverse set of ways to implement them (Cardinal et al. 2017). Generally, research on the employment of controls has been directed at understanding effective approaches to aligning workers' attitudes, intentions and behavior with an organization's goals and objectives. Next, the chapter presents the various approaches to classifying controls.

3.2.3 Formality of Controls

The actual procedures or practices used to implement controls can be viewed as either informal or formal. Informal controls are implemented by workers. They represent a shared set of beliefs and values among workers driven in part by their social relationships (Eisenhardt 1985; Ouchi 1979). Informal controls are often implicitly understood as a set of acceptable and unacceptable actions (Ouchi 1980). The consequences of violating them often include being expelled or ostracized from one's social group (Liu 2015). On the contrary, formal controls rely on explicitly stated rules or procedures that outline acceptable and unacceptable actions (Eisenhardt 1985; Kirsch 1997; Ouchi 1979). They are often driven by the management, and workers may or may not agree with them. In fact, workers often have little to no influence on determining formal controls. The consequences of violating a formal control involve the official actions by the organization.

Ideally, informal and formal controls should be aligned, but often they are not. It is possible for an employee to conform to a formal control and violate an informal control. Likewise, it is possible to conform to an informal control and violate a formal control. For example, workers who cross picket lines during an illegal strike might be violating an informal control while conforming to a formal control. These workers might keep their job but be expelled from their social group (i.e., union workers).

3.2.4 Control Systems

There are four types of control systems, i.e., configurations of multiple formal and informal controls. These include market, bureaucratic, clan, and integrative control systems (Cardinal et al. 2010). Market control systems are designed to focus on evaluating transaction outcomes such as the cost to perform a job. Market control systems do not rely heavily on either formal or informal control mechanisms. Bureaucratic control systems instead focus on specifying, monitoring, and evaluating the performance of workers (Ouchi and Price 1978). Bureaucratic control systems rely heavily on formal control mechanisms such as organizational rules, regulations, and procedures. Clan control systems emphasize aligning workers' motivations, beliefs, and values with those of the organization (Kirsch et al. 2010; Liu 2015). Clan control systems rely more on informal control mechanisms such as appealing to workers' personal pride or their identification with the organization. Finally, integrative control systems leverage both formal and informal control mechanisms (Cardinal et al. 2004; Jaworski and Kohli 1993; Sitkin and George 2005). For example, integrative control systems might employ formal controls such as rules and procedures along with informal controls such as appealing to workers' pride.

3.2.5 *Control Focus*

Controls can also be classified by the areas they are designed to influence: input, behavior, and output (Cardinal et al. 2017). Input controls focus on selecting the inputs (e.g., people and materials) that go into the work processes (Cardinal et al. 2010). This is often done by filtering out inputs that are seen as substandard. Typically, input controls are embedded throughout the hiring process of many organizations. For example, this would include requiring specific entrance exam scores or educational achievements before a person could be hired. Other examples include requiring potential suppliers to be certified before they can bid to provide manufacturers with raw materials. Input controls assume that if the inputs are of a certain quality it is more likely that the process will produce an acceptable output.

Behavior controls focus on aligning behaviors used to transform a set of resources such as labor and raw materials to a specific output such as the completion of a task or set of tasks. Behavior controls are directed at work processes needed to accomplish work (Robert 2016; Tiwana 2010). Behavior controls assume that if employees align their behavior to a predefined behavior or set of behaviors they are likely to perform a given task well (Dennis et al. 2012). Behavior controls include creating plans, defining work assignments, explicating work processes, and providing status reports on work (Piccoli and Ives 2003; Robert 2016). Behavior controls are effective when workers align their behavior to act in accordance with the established rules and procedures (Dennis et al. 2012; Robert 2016).

Output controls focus on influencing workers by holding them accountable to a predefined output metric (Choudhury and Sabherwal 2003; Kirsch 1997; Maruping et al. 2009). Output controls are directed at the final products or services produced and ignore the processes needed to accomplish the work. Output controls assume that if workers are held accountable for a predefined output they will align their behavior to achieve this output. Examples of output controls include paying factory workers for the number of correctly completed products rather than for the number of hours worked to complete the products. Output controls also include yearly, monthly, and quarterly goals for sales and production volumes.

Input, behavior, and output controls have advantages and disadvantages. In many cases, output controls can be very costly. This is because discovering that the final product is below standards means in many cases that resources that have been allocated were wasted. It is also costly in that any other task dependent on the final output is now held up. On the contrary, behavior controls allow for the continuous evaluation of work, which allows for problems to be identified and corrected sooner. Input controls are often the least costly when one considers the resources involved later in the transformation process, but this varies by industry. Input controls are often necessary but not sufficient to ensure successful output. The use of unqualified personnel is likely to lead to poor outputs, but the use of qualified personnel does not ensure high-quality outputs.

Input and output controls also have advantages. They do not require knowledge of the work process itself, nor do they require detailed planning to implement. This

is important in creative knowledge work, where the work processes are often not understood or cannot be seen. Hiring the most talented people and holding them accountable for what they produce and not how they produce it is an example of employing input and output controls. However, behavior controls do require knowledge of the work processes to create a predefined set of behavior standards. Behavior controls also require the ability to monitor the work processes. This can be problematic for creative knowledge work where work processes are less known and work is less visible.

3.2.6 Control Source

Who determines what controls are needed and how they should be implemented? This question speaks to the source of control. Sources of control include any entity that can impose controls. For example, in crowdsourcing, there are at least five sources of control: (1) platform providers, (2) requestors, (3) crowds, (4) sub-crowds, and (5) individuals within sub-crowds. Platforms provide the digital labor markets that connect workers to requestors who want to employ them. For example, Mechanical Turk and CrowdFlower are two popular digital platforms. Digital platforms can impose controls on crowd workers. Many platforms require crowd workers to maintain minimum performance standards. Requestors are another source of control. Requestors hire crowd workers and can employ controls to influence their behavior (Ye et al. 2017). Crowds themselves can exert control over their members. It is quite possible that controls can be exerted by multiple sources simultaneously, each with pros and cons. For example, Robert (2016) demonstrated that controls imposed by the group itself lead to better performance when compared to controls imposed by someone outside the group.

3.2.7 Control Unit of Analysis

Controls can be designed to influence organizations, groups, individuals within groups, and individuals. Controls directed at groups hold groups accountable rather than any individual within the group. For example, a group project completion date would be a group output control, whereas a task completion date for a specific individual would be an individual output control. This chapter differentiates between controls directed at individuals and controls directed at individuals within a group. Controls directed at individuals within a group are focused on aiding the collaborative work of the group, whereas controls directed at individuals who are not within a group are not focused on aiding collaborative group work. Therefore, controls directed at individuals within groups could be used to help promote macro-tasking, whereas controls directed at individuals outside of groups tend to be used to promote micro-tasking.

3.3 Scoping Literature Review

The authors of this chapter employed a scoping literature review to identify the various approaches to employing controls in crowdsourcing. The purpose of a scoping review is to rapidly map out the underpinnings of a research area (Mays et al. 2001). Scoping reviews provide an overview of a broader topic, whereas systematic reviews tend to have a narrow focus with an emphasis on depth (Peterson et al. 2017). The purpose of this scoping review was to survey the topic of controls in crowdsourcing and map out the various approaches used in the literature.

3.3.1 Literature Review Search

The literature review was conducted using Google Scholar. Google Scholar ranks articles by their relevance to the search topic and covers a wide and broad set of literature. This allowed the review to cut across several research areas covering controls in crowdsourcing. The search keywords were “controls” and “crowdsourcing” and the search was conducted in September 2018. The initial search identified 58,000 articles. The authors of this chapter evaluated article abstracts against the following inclusion and exclusion criteria.

Inclusion criteria. Studies were included if they (1) were empirical crowdsourcing studies and mentioned the use of controls and (2) were published in English-language journals/conferences.

Exclusion criteria. Studies were excluded if (1) they focused on types of controls that did not apply to the crowd or its members, (2) controls were in reference to variables such as age and gender (i.e., control variables), (3) they focused on control as an experimental procedure, or (4) they were nonempirical papers.

The literature search stopped at the first 370 articles identified by Google Scholar, for two reasons. First, articles beyond the first 300 became less and less relevant to the topic of control in crowdsourcing as outlined by the inclusion criteria. In fact, at the mark of 370, the articles met very few, if any, of the inclusion criteria. Second, the articles that were relevant did not add new knowledge to the scoping literature. In other words, the papers that were relevant employed controls no different from those in the papers already included in the review corpus.

The initial screening of the article abstracts produced 192 articles from the 370. Further analysis showed that 30 articles fell under exclusion criteria 1, 2, or 3, while 52 others fell under exclusion criterion 4, so they were dropped from the analysis. The remaining 110 articles met all inclusion criteria. Appendix 1 presents a summary table of the included articles.

Table 3.1 Publication venues, sources of control, and task type

Publication venues		Sources of control		Task type: macro versus micro	
Journals	63%	Requestor	92%	Micro	94.6%
Conferences	26%	Requestor and platform	5%	Macro	2.7%
Others	11%	Requestor and crowd members	3%	Macro and micro	2.7%

3.3.2 *Publication Venues*

The publication venues of the 110 included articles were as follows: 69 (63%) were published in journals, 29 (26%) were published in conferences, 8 (7%) were workshop papers, 3 (3%) were book chapters, and 1 (1%) was a research report. Although the journal and conference listings were diverse, many were published by ACM or IEEE (Table 3.1).

3.3.3 *Sources of Control*

Reviewers identified sources of control in each paper. Three sources of control were identified: platform, requestor, and crowd members (i.e., peers). Articles that employed platform controls relied on a predefined control embedded within the platform. An example of the use of a platform control would be to only include master turkers (MTurk crowd workers) in a study. The criteria used to determine who is or is not a master turker are set by the platform. The requestor was by far the most widely used source of control, employed in 101 (92%) papers; this was followed by the platform and requestor controls, used in 6 (5%) papers, then requestor and crowd members (peers) controls, used in 3 (3%; Table 3.1).

3.3.4 *Macro Versus Micro*

Reviewers determined whether the controls in each paper were focused on micro- or macro-tasks. Generally, studies that required participants to engage in simple standalone tasks without any need to coordinate with others were identified as micro, while studies that employed tasks that were not broken down and required coordinating with others were labeled as macro. The controls employed in crowdsourcing overwhelmingly focused on micro-tasking. One hundred four (94.6%) articles focused on micro-tasking while only 3 (2.7%) focused on macro-tasking. Three (2.7%) articles focused on both micro- and macro-tasking controls. See Table 3.1.

Table 3.2 Level of analysis and control type

Level of analysis		Control type			
Individuals	97%	Input	23%	Input and output	35%
Within groups or groups	3%	Behavioral	16%	Behavior and output	5%
		Output	83%	Input, behavior, and output	5%
Total	100%	Should not equal 100%		Total	45%

3.3.5 Level of Analysis

The paper findings on the level of analysis were consistent with those by Daniel et al. (2018). As stated by Daniel et al., “the quality and benefit of group work are still not fully studied and understood” (p. 29). Only 3 (3%) the articles focused on controls directed at individuals within groups, or groups, whereas 107 (97%) focused on controlling individuals (Table 3.2). This fully supports Daniel et al.’s additional conclusions that in failing to address issues of group control we also fail to fully leverage the potential of crowds.

3.3.6 Control Type

The authors of this chapter reviewed articles to determine the types of controls employed: input, behavior, or output, or any combination. Output controls were used the most, with 91 (83%) of the articles employing some type of output control (Table 3.2). Originally, the evaluation of crowd members’ output was done by humans; more recent work has shifted toward the use of advanced forms of artificial intelligence (AI). These approaches vary from relatively simple to more complex and are designed to better predict and evaluate worker outputs (e.g., Kajino et al. 2014). Yet, other approaches have sought to use both human and artificial intelligence systems (e.g., Haas et al. 2015).

Input controls and behavior controls were used less often than output controls. Input controls were used in 25 (23%) articles. The most common use of input controls was entrance tests to participate in the crowd work (Bozzon et al. 2013). Behavior controls were the least employed type of control, appearing in 18 (16%) articles. Types of behavior controls included real-time feedback on task performance, which allowed crowd workers to redo and improve their work, and design of better user interfaces to reduce error (e.g., Ashikawa et al. 2015; Gadiraju et al. 2015). See Table 3.2.

Nearly half of the articles (50, or 45%) employed more than one type of control. The most popular combination was input and output controls (39 articles, or 35%). This combination was typically employed by requiring an entrance test to participate in the work, then performing quality checks on the work performed (e.g., Eickhoff

and de Vries 2013; Hutton et al. 2012). Five articles (5%) employed both behavior and output controls, and 6 (5%) employed all three controls (Table 3.2).

3.3.7 Formality

The review found no evidence of informal controls. Because it was a scoping review, this does not mean that there was no use of informal controls but rather that they were rarely used when compared to formal controls.

3.3.8 Major Findings

Three major findings were derived from the literature review. Although the review also showed empirical evidence of other findings, the following insights represent the most consistent and generalizable results.

1. Crowdsourcing literature has focused primarily on the individual engaging in micro-tasking, with little attention directed at groups engaging in macro-tasking. As a result, we know very little about controls for macro-tasking involving groups.
2. The requestor has been the source of control and has relied heavily on output controls, with some efforts to leverage platform controls. On one hand, this approach does not require the requestor to have any knowledge of the work process. On the other hand, output controls alone are not enough to help the crowd manage and coordinate the work of its members. To accomplish this, the crowd itself must be leveraged as a source of control.
3. The literature on controls in crowdsourcing has focused mainly on formal controls. Yet informal controls can be as effective, if not more so, than formal controls (Kirsch et al. 2010). Informal controls also have the additional benefit of being more effective at promoting group cohesiveness.

3.4 Recommendations for Future Research

This section outlines a research agenda as a roadmap for future research by giving specific suggestions on how to shift toward the study of crowdsourcing controls for macro-tasking. Our research agenda is based on three assumptions:

1. Macro-tasks are not decomposed when assigned to a crowd; therefore, they require the crowd to decompose the task. In many cases, the tasks are not decomposable.
2. Macro-tasks require some degree of interaction and coordination among crowd members.

3. Macro-tasks require crowd members to undertake a diverse set of activities to accomplish their work. In other words, all crowd members do not perform the same task (i.e., little redundancy).

Given these assumptions and the gaps in the literature, this research agenda focuses on informal as well as formal controls for groups. The research agenda for formal controls not only includes input and output controls but also emphasizes the importance of behavior controls. To capture the effects of the group, this chapter conceptualizes crowds as a higher order structure that can exist on a given platform. Please see Fig. 3.1 for a visual depiction. Platforms are the digital technology that can host multiple crowds. In macro-tasking, crowds are groups of individuals working to achieve an overall common or shared goal. Crowds can be composed of multiple subgroups or sub-crowds. The term “sub-crowds” has been used by other scholars to represent smaller groups within the crowd (Malhotra and Majchrzak 2014). This chapter defines sub-crowds as crowd members who work independently to accomplish an objective that helps the crowd achieve its overall goal. Sub-crowds have boundaries in that there are members and nonmembers of sub-crowds. This boundary requirement applies even if membership is fluid. Sub-crowds can vary in size ranging from at least two crowd members. Macro-tasks that cannot be decomposed to micro-tasks are likely to be assigned to sub-crowds. Therefore, this chapter asserts that sub-crowd controls are a missing but vital component to understanding macro-tasking in crowds. In all, the research agenda’s focus on informal as well as formal

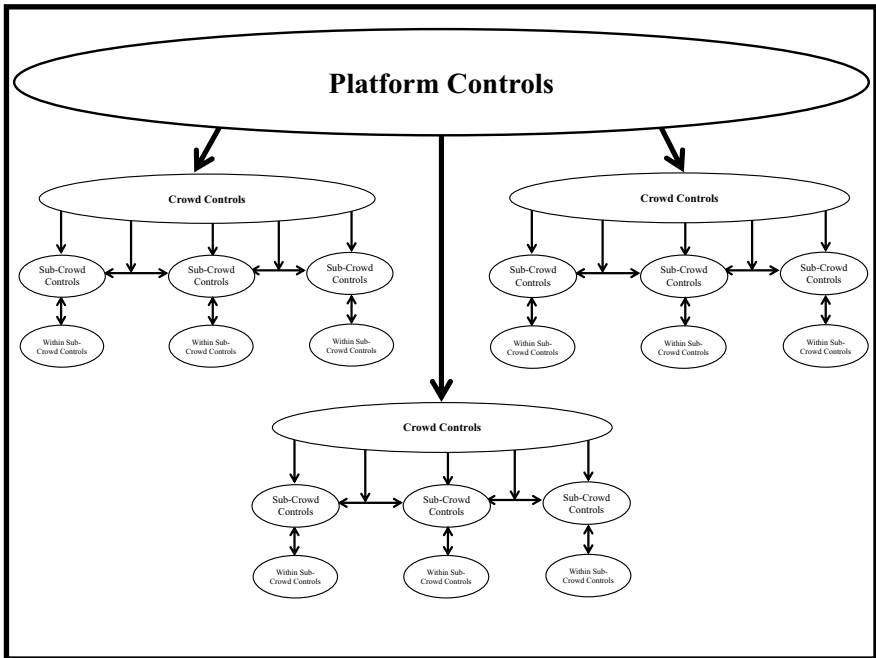


Fig. 3.1 Levels of crowdsourcing control

controls, the inclusion of crowds and sub-crowds as sources of control, and increased attention on behavior controls are expected to help address core shortcomings in the literature.

3.4.1 *DE-CoRe Control Framework*

To help identify the steps involved in the developmental of controls, this paper introduces the Defining, Evaluating, Correcting and Redefining (DE-CoRe) control framework. The DE-CoRe framework consists of four activities, listed next.

1. *Defining* involves developing and setting standard(s) that will be used later to compare against actual actions. These standards could refer to input, behavior, or output standards. Prescribed standards are the backbone of any control system. Standard setting for crowdsourcing input controls would focus on defining the selection criteria for potential crowd workers. For behavior control, it includes defining the behavior standards needed to perform the crowd work. Standard setting for output control would involve defining what constitutes a quality output.
2. *Evaluating* involves assessing the actual inputs, behaviors, and outputs against those prescribed standards. For input controls, this would involve evaluating potential crowd workers against the established selection criteria. Evaluation via behavior controls would involve comparing actual crowd worker behavior with the predefined behavior standard. Output control evaluation would determine whether the outputs produced met the predefined standard.
3. *Correcting*, if needed, involves identifying why and how inputs, behaviors, and outputs failed to meet the standards. This information provides feedback to explain what needs to be done differently to meet the prescribed standards. Correcting activity is distinct from the evaluation activity. Evaluation determines whether actions meet or fail to meet a predefined standard. Correcting activity focuses on why or how the actions failed to meet the predefined standard.
4. *Redefining*, if needed, is the final activity. For input control, this could entail changing the selection criteria. This might occur when new knowledge or skills are needed by crowd workers. In case of behavior controls, the need to redefine standards might be driven by new technology. For output control, quality standards can be redesigned based on new requirements.

In all, the DE-CoRe control framework provides a simple model to help organize and better communicate the research agenda presented in the next sections. Figure 3.2 depicts the developmental process and the iterative nature of the defining, evaluating, correcting, and redefining activities.

DE-CoRe Control Model

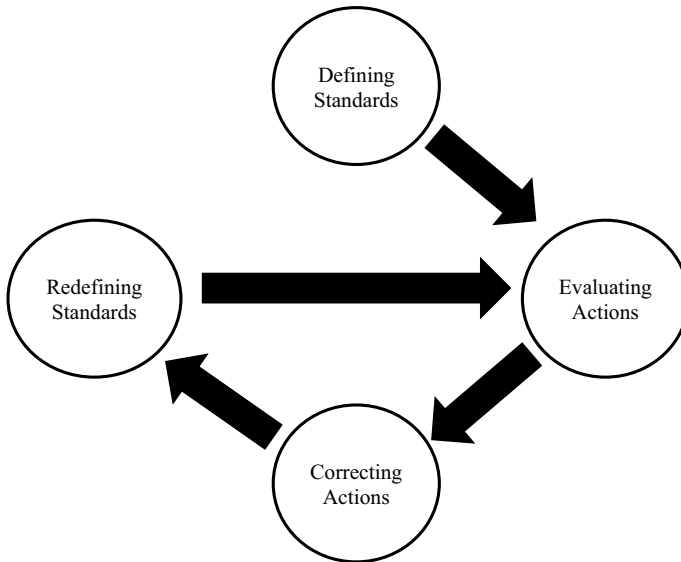


Fig. 3.2 DE-CoRe control framework

3.5 Formal Controls Research Agenda for Crowdsourcing Macro-tasking

Crowd: Input Controls

Research Question 1a: What are the most effective ways for crowds to employ input controls to promote crowdsourcing macro-tasks?

Definition. Crowd input controls are directed at the selection of the inputs (e.g., people and software) that go into the work processes of its sub-crowds. Crowd input controls ensure that the crowd inputs meet the predefined standards needed to support the achievement of the overall crowd's goals and objectives.

Examples. Examples include knowledge, skills, personality, and experience selection requirements, and minimum reputation scores.

Challenges. The problem of input control for macro-tasking in crowds is threefold: First, the set of knowledges and skills needed to complete macro-tasks might not be known because all the macro-task requirements might not be immediately identifiable. Second, the knowledge and skills needed might vary greatly depending on the task requirements of the assigned sub-crowd. This makes it difficult to determine whether one set of selection criteria should be used for all crowd workers or a different set of selection criteria should be used for each particular sub-crowd. For reasons

one and two, the crowd selection criteria should be more general, focusing on basic requirements for crowd workers. Finally, who should determine the selection criteria needed to employ the input controls: the requestor, the crowd, or the sub-crowds?

Design requirements

Defining. Systems must be able to help crowds determine the selection criteria for potential crowd workers. Such systems could allow crowds to leverage information from other crowds. For example, new crowds could use the work requirements from similar macro-tasks to determine the knowledge and skills needed by their crowd workers.

Evaluating. Going beyond filtering potential crowd workers by attributes, systems should be able to aid crowds in their decision-making process. A system might produce a list of recommended crowd workers based on the selection criteria. However, the system could go beyond this by rank ordering the list of crowd workers from most to least qualified. To promote diversity, the list could highlight the underrepresented minorities. To avoid problems of bias, the system could also alert the crowd when the selection criteria produce a list with no underrepresented minorities. Of course, what is and is not an underrepresented minority and whether a list should consider such factors is beyond the scope of this chapter.

Correcting. After crowd work has started, systems should be able to help crowds determine whether the selection criteria are being employed correctly. This would involve answering questions such as this: Are the selection criteria being ignored or incorrectly applied?

Redefining. Systems should support the redefining of selection criteria by using actual crowd worker performance. Crowds need answers to questions such as, “How predictive were the selection criteria in determining actual crowd worker performance across sub-crowds?” To this end, systems should produce reports that identify predictive selection criteria against actual performance data. Crowds could also leverage what they learned from the correcting activity to employ more effective selection criteria. For example, crowds might discover that their selection criteria were being ignored because they were ineffective.

Crowd: Behavior Controls

Research Question 1b: What are the most effective ways for crowds to employ behavior controls to promote crowdsourcing macro-tasks?

Definition. Crowd behavior controls direct the behavior of sub-crowds toward the achievement of the crowd’s goals and objectives. At the crowd level, behavior controls should be focused on ensuring effective interactions among sub-crowds. Therefore, crowd behavior controls should be directed at establishing standards to help govern how sub-crowds engage with one another.

Examples. Examples of behavior controls include sub-crowd status reports and lists of completed or uncompleted work.

Challenges. The biggest challenge with regard to crowd behavior controls is to determine how much autonomy should be afforded to sub-crowds. This is particularly problematic when many of the work requirements are not initially known. Therefore, crowd behavior controls should foster cooperation among sub-crowds while providing them with the needed autonomy to develop their own behavior controls after work requirements become known. Specifically, crowd behavior controls should be directed at creating standards for communication and interaction among sub-crowds. Crowds should pay attention to work dependencies that require hand-offs among sub-crowds. Crowd behavior controls should be developed to avoid or resolve problems that slow or hinder the transfer of work among sub-crowds.

Design requirements

Defining. Going beyond basic communication requirements, systems should help identify work dependencies across sub-crowds. This would help crowds understand the requirements needed to ensure effective handoffs of work among sub-crowds.

Evaluating. To help with evaluation, systems should support the creation of digital boundary objects. Boundary objects are artifacts employed to track activity across group boundaries (Star and Griesemer 1989). Within crowd work, digital boundary objects are electronic artifacts employed to track work across multiple sub-crowds. Digital boundary objects are vital to assisting crowds in monitoring and tracking the work of sub-crowds. Although boundary objects are common to most work, such as “to-do lists,” some boundary objects are context-dependent. Therefore, systems must have the flexibility to allow crowds to construct their own digital boundary objects when needed.

Correcting. To support the correcting activity, systems must produce work reports that highlight where sub-crowds went wrong and how to correct their actions. These reports could focus on identifying the sub-crowd that failed to meet requirements. This would answer questions such as this: Which sub-crowd failed to report what information when?

Redefining. Systems should allow crowds to redefine work standards when needed. After employing behavior controls, crowds might realize that their current reporting requirements are: (1) simply not enough to promote effective communication and interaction or (2) too cumbersome for sub-crowds to follow. Systems that could help to diagnose either problem and allow crowds to leverage this information in redefining their behavior standards would be invaluable.

Crowd: Output Controls

Research Question 1c: What are the most effective ways for crowds to employ crowd output controls to promote crowdsourcing macro-tasks?

Definition. Crowd output controls ensure that sub-crowd outputs meet the crowd’s predefined output standards or set of standards. Crowd output controls are used to hold sub-crowds accountable by making it clear what is and is not an acceptable output. Crowd output controls should ensure that sub-crowds are supporting the

crowd's overall goal and not engaging in suboptimization at the expense of other sub-crowds.

Examples. Examples of this include sub-crowd goals for completed work, sub-crowd goals for correctly completed work, and deadlines for completed work.

Challenges. The interdependent nature of macro-tasking across sub-crowds presents several challenges. First, macro-tasking often requires output from one group to be used by other groups. Such dependencies must be identified before group output controls can be designed and employed. Crowds would also need to build consensus among groups on what such output controls should be when such dependencies exist. The second and related issue is that group output controls must align across groups. An example of misalignment is when one group in the crowd is being evaluated on quantity but the group receiving the output is more concerned about quality. The group producing the output might ignore issues related to quality to achieve more quantity. Yet this would be all for naught, because the output would be useless to the receiving group if the quality was not acceptable.

Design requirements

Defining. Systems must allow crowds to define output standards by identifying quality criteria and assigning value weights to such criteria. Systems with advanced capabilities might provide cost-benefit calculations. This would allow crowds to understand trade-off between decisions regarding quality and quantity. Going beyond this requirement, advanced systems would need to help crowds deal with issues related to the task interdependence among sub-crowds. To avoid problems related to suboptimizing, systems should aid in the identification of work dependencies.

Evaluating. Systems should provide tools to assess or help assess the quality of crowd outputs. These tools could be designed to help crowds manually assess quality or be completely automated.

Correcting. If needed, systems should produce reports that help crowd workers understand why and how they are failing to meet output standards. For example, are the failures related to quantity or quality or both? Should sub-crowds focus on doing less but better?

Redefining. Similar to behavior control, systems should allow crowds to redefine their output standards. Systems could help leverage the information generated in the correction activity. If many sub-crowds are failing to meet deadlines, maybe the deadlines should be changed. If the sub-crowds are meeting output standards regarding quantity easily, maybe such output standards should be increased.

Sub-crowd: Input Controls

Research Question 2a: What are the most effective ways for sub-crowds to employ input controls to promote crowdsourcing macro-tasks?

Definition. Sub-crowd input controls focus on the inputs that go into the sub-crowd's work processes. Like crowd input controls, sub-crowd input controls would primarily

focus on selection criteria for membership. However, they could also include the selection of software or other collaborative tools. Sub-crowd input controls offer another opportunity to employ controls that help promote macro-tasking.

Examples. Examples include knowledge, skills, personality, and experience selection requirements *over and above those required by the crowd*, and minimum reputation scores *over and above those required by the crowd*.

Challenges. Several issues arise when considering sub-crowd input controls. First, it is important to determine what additional selection criteria might be needed for sub-crowd membership above those required for crowd membership. This entails determining the sets of knowledge and skills needed to complete the sub-crowd's work. This could also include increasing the required scores needed on the crowd's selection criteria. For example, sub-crowds might require higher technical skills depending on the nature of their work. Second, it would be necessary to determine whether the sub-crowds' selection criteria superseded the crowd's selection criteria or vice versa. It would also be important to know whether sub-crowds could completely bypass the crowd's selection criteria. For example, could sub-crowds select individuals who had been rejected by the crowd? This is important because sub-crowds might have the opportunity to hire unqualified crowd workers and provide training that would eventually make them qualified. Sub-crowds could evaluate the performance of such crowd workers after a trial period to determine whether they should be retained.

Design requirements

Defining. In addition to the design requirements outlined for defining crowd input controls, systems must be able to help sub-crowds identify any conflicts between their and the crowd's input controls. For example, such systems would need to identify potential conflicts between the crowd and sub-crowd selection criteria.

Evaluating. The evaluating requirements for sub-crowd input controls should be similar to those for crowd input controls.

Correcting. In addition to the correcting requirements outlined for crowd input controls, systems should be better designed to provide more flexibility in allowing sub-crowds to override their selection criteria. These systems should require an acknowledgment and a thorough explanation as to why the selection criteria are being ignored. Unlike the crowd selection criteria, which are likely to be more general and stable, the sub-crowd selection criteria are likely to be more specific and dynamic. Sub-crowd selection criteria are likely to change rapidly as work requirements become clearer and work progresses. Therefore, sub-crowds might not have the luxury to wait for the redefining activities to change selection criteria. In fact, depending on the work duration, sub-crowds might disband before they ever reach the redefining activity.

Redefining. Processes for redefining sub-crowd input control requirements should be similar to those for the crowd input control requirements.

Sub-crowd: Behavior Controls

Research Question 2b: What are the most effective ways for sub-crowds to employ behavior controls to promote crowdsourcing macro-tasks?

Definition. Sub-crowd behavior controls focus on aligning the behaviors of the sub-crowd workers with the behaviors needed to achieve the sub-crowd's goals and objectives. Although sub-crowd behavior controls are concerned with effective interactions among sub-crowd workers, they also specify work standards needed to accomplish work tasks. Therefore, when compared to crowd behavior controls, sub-crowd behavior controls should be more detailed and task specific.

Examples. Examples of sub-crowd behavior controls include work instructions, crowd worker status reports, lists of crowd workers' completed or uncompleted work, shared calendars, and work assignment spreadsheets.

Challenges. It would be difficult to assemble sub-crowds with no common work history and expect them to work together to develop behavior controls without any guidance. In other words, newly formed sub-crowds would need behavior controls to begin to work together to develop behavior controls. This chapter proposes conceptualizing behavior controls as those employed before and after the sub-crowd workers develop knowledge of their work requirements. To address this challenge, this chapter introduces Layer 1 and Layer 2 behavior controls.

Layer 1 behavior controls are standards directed at helping the sub-crowd determine the work requirements. Layer 1 behavior controls can be imposed by the crowd or quickly agreed upon by the sub-crowd. In the first approach, the crowd could dictate initial basic sub-crowd behavior controls. This approach could be referred to as the template approach to behavior controls. Templated behavior controls should be generic and light and apply broadly to any sub-crowd. These template behavior controls can be viewed as basic rules of engagement for crowd workers. Sub-crowds could then develop their own behavior controls later when work requirements became clearer. In the second approach, sub-crowds could engage in swift planning via a sub-crowd charter. A sub-crowd charter is a document that outlines the sub-crowd's objectives and communication protocols, and crowd workers' basic roles and responsibilities. Sub-crowds could add or remove requirements to their charter as work progressed. The differences between the first and second approaches to developing Layer 1 behavior controls are a matter of degree. Simply put, the two approaches vary on the degree to which the crowd or the sub-crowd has an initial influence on the Layer 1 behavior controls. Therefore, the third approach would be for the crowd to provide a template in line with the sub-crowd character and enlist the sub-crowd to decide which aspects to keep and which to remove.

Layer 2 behavior controls are directed at defining standard behaviors needed to perform work. There are two approaches to developing Layer 2 behavior controls. The first approach is to have the sub-crowd workers determine them as their work requirements become clear. Layer 2 behavior controls provide instructions on how crowd workers should accomplish their job. The degree of detail associated with the instructions depends on the effort and time needed to specify such detail. Ideally, sub-crowds should weigh the benefits associated with such specification against the time

and effort needed. The second approach is to provide sub-crowds with work standards already developed based on best work practices. Like the template approach to Layer 1, these best work practices would be generic and light and apply broadly. However, they could also be very detailed if the new work requirements were similar to previous work requirements from another sub-crowd or crowd. Like the two approaches to Layer 1, the two approaches to developing Layer 2 can also be combined. Therefore, the third approach would involve the sub-crowd starting with a template based on best practices and customizing it to the sub-crowd's needs.

In either case, Layer 1 behavior controls should be removed or changed if they prevent the actual work from being accomplished. At the same time, Layer 1 behavior controls might be sufficient to accomplish the sub-crowd work; if this occurs, there is no need to define Layer 2 behavior controls.

Design requirements

Defining. Going beyond basic communication requirements, systems should provide tools to help sub-crowds break down, structure, assign, and aggregate crowd work. Such systems could provide digital workflow diagrams, shared calendars, and work assignment spreadsheets.

Evaluating. To help with evaluation, systems should afford the design of digital artifacts such as to-do lists and crowd worker status reports. These digital artifacts would be similar in concept to the digital boundary objectives but different in at least two ways: (1) these artifacts would not be designed to be used by other sub-crowds and (2) they would be focused on evaluating the behavior of sub-crowd workers rather than the sub-crowd itself.

Correcting. Systems must produce work reports that show where sub-crowd workers went wrong and how to correct their actions. These reports should be more detailed than those produced for crowds.

Redefining. After employing behavior controls, sub-crowds might realize that they were: (1) ineffective even when followed correctly or (2) too difficult for crowd workers to follow correctly. In either case, sub-crowds would have to redefined work standards. Ideally, sub-crowds should be able to leverage the same system capabilities used in the defining phase. However, new system capabilities might be needed when new work standards are vastly different.

Sub-crowd: Output Controls

Research Question 2c: What are the most effective ways for sub-crowds to employ output controls to promote crowdsourcing macro-tasks?

Definition. Sub-crowd output controls ensure that the output of crowd workers in a sub-crowd meets the sub-crowd's predefined output standards or set of standards. Sub-crowd output controls hold crowd workers accountable to a predefined outcome or set of outcomes identified as vital to achieving the sub-crowd's overall goals and objectives. **Note:** Output controls are likely to be very important to sub-crowds engaging in complex and creative macro-tasks. In such cases, output controls are often

preferred over behavior controls. This is because specifying detailed instructions for complex and creative tasks is very difficult. In addition, creative work is often not visible; as such it is hard to monitor and track the progress of creative work.

Examples. Examples of sub-crowd output controls include crowd worker lists of correctly completed tasks, the number of completed tasks, and due dates for completed tasks.

Challenges. The degree of task heterogeneity and its corresponding output control heterogeneity is likely to be a major challenge. The tasks of crowd workers within a given sub-crowd are likely to be related and interdependent—related in that all tasks performed by crowd workers in the same sub-crowd would be directed at achieving a common goal, and interdependent in that the output of every crowd worker within a sub-crowd would need to be aggregated before the sub-crowd could achieve its goals.

Yet, crowd workers' tasks are likely to be different. Task heterogeneity might require a diverse set of output controls among crowd workers within the same sub-crowd. For example, for one task, the quantity might be far more important than quality. But for another task, deadlines might be the most important factor. Finding a way to harmonize the output controls needed to avoid conflicts within a sub-crowd is likely to be problematic. In addition, incompatible output controls are likely to lead to low sub-crowd cohesion.

Design requirements

Defining. In addition to the design requirements outlined for defining crowd output controls, systems supporting sub-crowds should place more emphasis on issues related to task heterogeneity. More specifically, how can such systems help sub-crowds harmonize output controls to avoid controls conflicting with one another?

Evaluating. Systems should provide tools to assess or help assess the quality of individual crowd workers. In addition, such systems should be able to evaluate small groups of crowd workers who perform a similar task, yet be flexible enough to evaluate individual crowd workers across a wide range of tasks.

Correcting. For correcting sub-crowd output controls, systems should be able to provide detailed reports on a range of tasks.

Redefining. Like crowd output controls, systems should allow sub-crowds to redefine their output standards.

Table 3.3 summarizes the formal controls research agenda.

3.6 Informal Controls Research Agenda for Crowdsourcing Macro-tasks

Research Question 3: What are the most effective ways to promote informal controls in crowds for macro-tasking in crowdsourcing?

Table 3.3 Formal controls and DE-CoRe design objectives

DE-CoRe design objectives		
Control	Design objectives	Exemplars
Input control RQs: 1a and 2a	Defining input standards <ul style="list-style-type: none"> • Selection standards <ul style="list-style-type: none"> – Identify knowledge and skills Evaluating inputs <ul style="list-style-type: none"> • Select qualified crowd workers • Qualify crowd workers <ul style="list-style-type: none"> – Train – Test Correcting inputs <ul style="list-style-type: none"> • Detailed work reports Redefining input standards <ul style="list-style-type: none"> • Revising selection standards 	Li et al. (2014) put forth a crowd targeting framework designed to automatically discover the needed crowd worker skills for a given task and target the most qualified crowd workers based on this skill set
Behavior control RQs: 1b and 2b	Defining behavior standards <ul style="list-style-type: none"> • Break down crowd work • Structure crowd work • Assign crowd work • Aggregate crowd work Evaluating behavior <ul style="list-style-type: none"> • Monitor crowd work • Assess crowd work Correcting behavior <ul style="list-style-type: none"> • Detailed work reports Redefining behavior standards <ul style="list-style-type: none"> • Break down crowd work • Structure crowd work • Assign crowd work • Aggregate crowd work 	Schmitz and Lykourantzou (2018) designed and empirically tested an online algorithm that engages in the structuring and scheduling of work to accomplish macro-tasks
Output control RQs: 1c and 2c	Defining output standards <ul style="list-style-type: none"> • Identify quality criteria • Assign value weights on criteria Evaluating output <ul style="list-style-type: none"> • Manual assessment tools • Automated assessment tools Correcting behavior <ul style="list-style-type: none"> • Detailed work reports Redefining output standards <ul style="list-style-type: none"> • Identify new quality criteria • Assign new value weights to criteria 	Oleson et al. (2011) offered a novel approach to assessing output quality by proposing new ways to develop gold standards used to assess crowd worker outputs

Many of the challenges and design requirements for informal controls are similar to those of formal controls. The biggest difference is the role that social relationships play in the employment of informal controls. Generally, informal control is a type of social control exerted by members of the collective. Informal controls influence actions by exerting normative peer pressure on crowd workers. A more specific definition of informal controls can be derived from Kirsch et al. (2010). According to Kirsch et al., informal controls are exerted when shared norms, values, beliefs, and vision influence the behaviors of the collective. This is consistent with literature identifying the need to facilitate social bonds, identification, and common values among members of a collective to help establish and strengthen informal controls (Weibel et al. 2016). However, social bonds, identification, and common values are normally associated with groups with a long history of working together (Robert et al. 2008).

Therefore, the biggest challenge associated with informal controls relative to formal controls is determining how crowd workers with little history can develop the social bonds, identification, and common values needed to employ informal controls. In this section, the discussion on informal controls is focused on addressing this issue only. However, some of the same challenges and design requirements identified in the discussion on formal controls are also applicable. In addition, this chapter acknowledges that depending on the task duration and task complexity, crowd workers may or may not have an opportunity or a need for informal controls. Yet, without informal controls, macro-tasking complex and creative work is likely to be difficult. Consequentially, informal control is likely to be difficult to establish but nonetheless very important in the crowdsourcing of macro-tasks. Next are several approaches to promoting informal controls in crowdsourcing macro-tasks. They are summarized in Table 3.4.

One approach is to understand how to help crowds build common norms, values, beliefs, and vision through the promotion of a shared identity. Research has shown that a shared identity can facilitate the establishment of common norms, values, beliefs, and vision (Chatman 2010; Robert 2016). Windeler et al. (2015) provided an example of how this approach could be operationalized. They studied ways to reduce conflict and promote a shared understanding and ultimately improve performance in online teams. They designed a system that provided one set of teams with profiles of each team member that only listed similar attributes among team members. This was done to promote perceptions of similarity—a shared or common identity among team members. Another set of teams received no such information regarding their similarities. The online teams that received the similarity information experienced

Table 3.4 Informal controls and design objectives

Informal control mechanism	Design objectives	Examples
Identification	Perceived similarity	Windeler et al. (2015)
Shared norms and values	Socialization/onboarding	Homan et al. (2007)
Identification, shared norms, and values	Familiarity	Salehi et al. (2017)

less conflict, had a better shared understanding, and performed better as a team. A similar approach could be used in crowdsourcing. Questions like how to best promote similarities or which similarities to promote still need to be addressed. Nonetheless, designing crowdsourcing systems to promote similarities among crowds or sub-crowds holds much potential.

Another approach is helping crowds establish shared work norms and values. In traditional organizations, new employees go through a socialization process that both introduces and facilitates preexisting shared norms, values, beliefs, and vision (Turner and Makhija 2006). Organizations often leverage orientation and training programs to help establish prototype norms, values, and beliefs. Similar approaches have been done in groups. For example, Homan et al. (2007) conducted a lab study and found that teams trained to value diversity were able to establish norms that led them to better leverage diversity to perform better. Crowdsourcing systems can be designed to not only train crowd workers but also orient workers to a specific crowd climate or culture. This could be done by building crowdsourcing systems that walk crowds or sub-crowds through series of group-building exercises. Although there are many unanswered questions related to finding effective team-building exercises and designing such a crowdsourcing system, this avenue holds the potential to promote informal controls.

Another approach to promoting informal controls is to select crowd workers who already have shared norms, values, beliefs, and vision. This could be accomplished by selecting crowd workers who worked together in the past. For example, a crowdsourcing system could be programmed to select crowd workers from a GitHub project. This system could be designed to assess the success of a group of crowd workers based on a specific metric. Then the system could invite all crowd workers who participated in a specific project or part of the project. These crowd workers would likely have been indoctrinated into a system of shared norms, values, beliefs, and vision. Salehi et al. (2017) provided an example of this approach. Their systems selected crowd workers based on whether they were familiar with one another. Familiarity is a strong predictor of shared norms, values, beliefs, and vision. By selecting specific online communities like GitHub, organizations could ensure they hire crowd workers who are competent in a specified domain. Questions about which parameters to use to select crowd workers along with the actual design of such systems needed to operationalize the selection criteria are important issues to be addressed.

3.7 Future Research and Limitations

The next section presents several limitations as well as future research opportunities. While these areas complement and overlap the research areas identified and discussed earlier in the chapter, these areas could themselves constitute their own research agenda. Although they could not be sufficiently discussed in detail in this chapter, they are important areas that should be acknowledged.

3.7.1 Meta-control Theory

To accommodate the use of multiple types of control inherent in the crowdsourcing of macro-controls, this chapter introduces the meta-control theory. Meta-control theory focuses on comprehending the impacts of controls on controls. Meta-control theory is concerned with understanding how controls reinforce or undermine one another. The goal of meta-control theory is to avoid controls conflicting with or undermining one another. Meta-control theory also recognizes that controls must be dynamically managed throughout their use. Meta-control theory acknowledges that controls make up a complex system that might not lead to linear, well-understood effects but instead could lead to nonlinear effects that are difficult to understand. Understanding how to ensure that controls align across levels of analysis is one example of meta-control theory.

The theoretical development and empirical validation of the study of how controls impact controls could significantly contribute to control theory in general as well as its specific application to crowdsourcing. Yet, we have not begun to scratch the surface in this area. Although we have empirical examples of the use of multiple controls, little theory or reasoning has been offered as to why these particular controls were chosen or how they are expected to align with one another or, better yet, when they are expected not to align with one another. This is almost certainly a result of the micro-tasking nature of most crowdsourcing work. Nonetheless, as we move toward macro-tasks, meta-control theory, or the study of how controls impact controls, is becoming increasingly important.

3.7.2 Temporal Effects on Control

Generally, things change over time. This is not surprising or profound—the impact and importance of time have been increasingly recognized by many HCI/CSCW scholars and others (You et al. 2015). Yet no studies of control examine the impact of time. At this stage, the evidence of the importance of time on controls is more anecdotal than scientifically verifiable. For example, platform companies like Uber update their controls based on dimensions such as time. For instance, by implementing surge pricing, Uber charges higher driving fares during peak demand times.

A less popular example of the impact of time on the effectiveness of control relates to Uber’s driver assignment algorithm. Uber’s driver assignment is a type of behavior control the company imposes on drivers. However, many drivers learn how Uber’s algorithm assigns which drivers to which routes. Drivers then attempt to manipulate their assignment to more lucrative routes. Uber responds by changing the assignment algorithm to prevent such manipulation. Hence, over time Uber’s behavior control has become less effective. A more systematic research agenda might not only investigate how time impacts the effectiveness of controls but why, when,

and how. What is certain is that we know little if any with regard to the impact of time on the effectiveness of controls in crowdsourcing.

3.7.3 Artificial Intelligence Control Systems

The use of artificial intelligence (AI) to control workers is becoming popular in many industries. AI—the ability of a computer system to sense, reason, and respond—holds many potential uses for controlling crowd workers for macro-tasking. Artificial intelligence control systems (AICS) are intelligent computer systems that *seek to align and dynamically realign workers' actions to predefined standards to achieve a set of goals and objectives*. AICS can dynamically evaluate, correct, and redefine controls in real time. AICS can be used as input, behavior, and output controls. There are several examples of researchers employing automated quality assessments (Hoßfeld and Keimel 2014) or automating work processes (Schmitz and Lykourantzou 2018). However, these systems fall far short of employing the full capabilities of AICS currently used in many digital platforms (i.e., Uber and Upwork). Future HCI/CSCW research needs to explore both the development and implications of AICS in crowdsourcing.

3.8 Conclusions

The conditions needed to design effective controls for micro-tasks represent an approach to control that is typical of the Industrial Age. But as crowd work becomes increasingly more complex, interdependent, and less decomposable, focusing more on innovation and learning than performing, HCI scholars must ask ourselves how we can design controls that better meet the demands of macro-tasking. The need to rethink controls for new ways of working is not a particularly new problem, nor is it confined to HCI scholars examining crowdsourcing. Organizational scholars have warned of the need for dramatic changes in our approaches to organizing and they have decried the lack of progress toward newer approaches to designing controls (Cardinal et al. 2010). As such, this chapter should help organizational scholars begin to rethink the design of controls in traditional organizational settings.

Acknowledgements This book chapter was supported in part by the National Science Foundation [grant CHS-1617820].

Appendix 1

	Publication type	Source	Task	Level	Control focus		
					Input	Behavior	Output
Aker, A., El-Haj, M., Albakour, M. D., & Kruschwitz, U. (2012). Assessing Crowdsourcing Quality through Objective Tasks. In Proceedings of the Eighth International Conference on Language Resources and Evaluation (pp. 14561461).	Conference	Requestor	Micro	Individuals			X
Ashikawa, M., Kawamura, T., & Ohsuga, A. (2015, December). Deployment of private crowdsourcing system with quality control methods. In Web Intelligence and Intelligent Agent Technology (WI-IAT), 2015 IEEE/WIC/ACM International Conference on (Vol. 1, pp. 9–16). IEEE.	Conference	Platform and requestor	Micro	Individuals	X	X	X

(continued)

(continued)

	Publication type	Source	Task	Level	Control focus		
					Input	Behavior	Output
Baba, Y., & Kashima, H. (2013, August). Statistical quality estimation for general crowdsourcing tasks. In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 554-562). ACM.	Conference	Requestor and peer	Micro	Individuals			X
Baba, Y., Kashima, H., Kinoshita, K., Yamaguchi, G., & Akiyoshi, Y. (2013, June). Leveraging Crowdsourcing to Detect Improper Tasks in Crowdsourcing Marketplaces. In Twenty-Fifth Innovative Applications of Artificial Intelligence Conference, pp. 1487- 1492.	Conference	Requestor	Micro	Individuals			X

(continued)

(continued)

	Publication type	Source	Task	Level	Control focus	
					Input	Behavior
Baba, Y., Kashima, H., Kinoshita, K., Yamaguchi, G., & Akiyoshi, Y. (2014). Leveraging non-expert crowdsourcing workers for improper task detection in crowdsourcing marketplaces. <i>Expert Systems with Applications</i> , 41 (6), 26782687.	Journal	Requestor	Micro	Individuals		X
Bell, S., & Bala, K. (2015). Learning visual similarity for product design with convolutional neural networks. <i>ACM Transactions on Graphics (TOG)</i> , 34(4), 98.	Journal	Requestor	Micro	Individuals	X	X

(continued)

(continued)

	Publication type	Source	Task	Level	Control focus		
					Input	Behavior	Output
Bontcheva, K., Roberts, I., Derczynski, L., & Rout, D. (2014). The GATE crowdsourcing plugin: Crowdsourcing annotated corpora made easy. In Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics (pp. 97–100).	Conference	Requestor	Micro	Individuals			X
Bozzon, A., Brambilla, M., Ceri, S., & Mauri, A. (2013, May). Reactive crowdsourcing. In Proceedings of the 22nd international conference on World Wide Web (pp. 153–164). ACM.	Conference	Requestor	Micro	Individuals	X	X	X

(continued)

	Publication type	Source	Task	Level	Control focus		
					Input	Behavior	Output
(continued)	Conference	Requestor	Micro/Macro	Individuals or individuals within groups	X	X	X
Bozzon, A., Brambilla, M., Ceri, S., Mauri, A., & Volonteri, R. (2014, July). Pattern-based specification of crowdsourcing applications. In International Conference on Web Engineering (pp. 218–235). Springer, Cham.	Conference	Requestor	Micro	Individuals			X
Bragg, J., & Weld, D. S. (2013, November). Crowdsourcing multi-label classification for taxonomy creation. In First AAAI conference on human computation and crowdsourcing.	Journal	Requestor	Micro	Individuals			X
Causser, T., Tonra, J., & Wallace, V. (2012). Transcription maximized; expense minimized? Crowdsourcing and editing the collected works of Jeremy Bentham. <i>Literary and Linguistic Computing</i> , 27(2), 119–137.	Journal	Requestor	Micro	Individuals			X

(continued)

(continued)

	Publication type	Source	Task	Level	Control focus		
					Input	Behavior	Output
Chang, D., Chen, C. H., & Lee, K. M. (2014). A crowdsourcing development approach based on a neuro-fuzzy network for creating innovative product concepts. <i>Neurocomputing</i> , 142, 60–72.	Journal	Requestor	Micro	Individuals			
Chen, Z., Fu, R., Zhao, Z., Liu, Z., Xia, L., Chen, L., ... & Zhang, C. J. (2014). gMission: A general spatial crowdsourcing platform. <i>Proceedings of the VLDB Endowment</i> , 7(13), 1629–1632.	Conference	Requestor	Micro	Individuals			X
Cheng, J., Teevan, J., Iqbal, S. T., & Bernstein, M. S. (2015, April). Break it down: A comparison of macro-and microtasks. In <i>Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems</i> (pp. 4061–4064). ACM.	Conference	Requestor	Micro	Individuals	X		X

(continued)

(continued)

	Publication type	Source	Task	Level	Control focus	
					Input	Output
Chiu, C. M., Liang, T. P., & Turban, E. (2014). What can crowdsourcing do for decision support?. <i>Decision Support Systems</i> , 65, 40–49.	Journal	Requestor	Micro	Individuals		X
Chung, M. J. Y., Forbes, M., Cakmak, M., & Rao, R. P. (2014, May). Accelerating imitation learning through crowdsourcing. In <i>ICRA</i> (pp. 4777–4784).	Conference	Requestor	Micro	Individuals		X
Dai, P., Lin, C. H., & Weld, D. S. (2013). Pomdp-based control of workflows for crowdsourcing. <i>Artificial Intelligence</i> , 202, 52–85.	Journal	Requestor	Micro	Individuals		X

(continued)

	Publication type	Source	Task	Level	Control focus		
					Input	Behavior	Output
(continued)	Conference	Requestor	Micro	Individuals		X	X
Dai, P., Rzeszotarski, J. M., Paritosh, P., & Chi, E. H. (2015, February). And now for something completely different: Improving crowdsourcing workflows with micro-diversions. In Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (pp. 628–638). ACM.	Conference	Requestor	Micro	Individuals			X

(continued)

(continued)

	Publication type	Source	Task	Level	Control focus		
					Input	Behavior	Output
Deng, J., Krause, J., & Fei-Fei, L. (2013). Fine-grained crowdsourcing for fine-grained recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 580–587).	Conference	Requestor	Micro	Individuals			X
Difallah, D. E., Demartini, G., & Cudre-Mauroux, P. (2012, April). Mechanical Cheat: Spamming Schemes and Adversarial Techniques on Crowdsourcing Platforms. In CrowdSearch 2010 Workshop at WWW 2012 (pp. 26–30).	Conference	Requestor	Micro	Individuals			X
Duan, L., Oyama, S., Sato, H., & Kurihara, M. (2014). Separate or joint? Estimation of multiple labels from crowdsourced annotations. Expert Systems with Applications, 41(13), 5723–5732.	Journal	Requestor	Micro	Individuals			X

(continued)

(continued)

	Publication type	Source	Task	Level	Control focus		
					Input	Behavior	Output
Eickhoff, C., & de Vries, A. (2011, February). How crowdsourcable is your task. In Proceedings of the workshop on crowdsourcing for search and data mining (CSDM) at the fourth ACM international conference on web search and data mining (WSDM) (pp. 11–14).	Conference	Requestor	Micro	Individuals			X
Eickhoff, C., & de Vries, A. P. (2013). Increasing cheat robustness of crowdsourcing tasks. Information retrieval, 16(2), 121–137.	Journal	Requestor	Micro	Individuals	X		X
Fan, J., Li, G., Ooi, B. C., Tan, K. L., & Feng, J. (2015, May). icrowd: An adaptive crowdsourcing framework. In Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data (pp. 1015–1030). ACM.	Conference	Requestor	Micro	Individuals			X

(continued)

(continued)

	Publication type	Source	Task	Level	Control focus		
					Input	Behavior	Output
Fang, Y., Sun, H., Li, G., Zhang, R., & Huai, J. (2016, April). Effective result inference for context-sensitive tasks in crowdsourcing. In International Conference on Database Systems for Advanced Applications (pp. 33–48). Springer, Cham.	Conference	Requestor	Micro/Macro	Individuals			X
Filatova, E. (2012, May). Irony and Sarcasm: Corpus Generation and Analysis Using Crowdsourcing. Proceedings of the Ninth International Conference on Language Resources and Evaluation (pp. 392–398).	Conference	Requestor	Micro	Individuals			X

(continued)

	Publication type	Source	Task	Level	Control focus	
					Input	Output
(continued)						
Finin, T., Murmane, W., Karandikar, A., Keller, N., Martineau, J., & Dredze, M. (2010, June). Annotating named entities in Twitter data with crowdsourcing. In Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk (pp. 80–88). Association for Computational Linguistics.	Workshop	Requestor	Micro	Individuals		X
Foncubierta Rodriguez, A., & Muller, H. (2012, October). Ground truth generation in medical imaging: a crowdsourcing based iterative approach. In Proceedings of the ACM multimedia 2012 workshop on Crowdsourcing for multimedia (pp. 9–14). ACM.	Conference	Requestor	Micro	Individuals		X

(continued)

(continued)

	Publication type	Source	Task	Level	Control focus		
					Input	Behavior	Output
Franklin, M. J., Kossmann, D., Kraska, T., Ramesh, S., & Xin, R. (2011, June). CrowdDB: answering queries with crowdsourcing. In Proceedings of the 2011 ACM SIGMOD International Conference on Management of data (pp. 6172). ACM.	Conference	Requestor	Micro	Individuals			X
Fu, W. T., & Liao, V. (2011, March). Crowdsourcing quality control of online information: a quality-based cascade model. In International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction (pp. 147–154). Springer, Berlin, Heidelberg.	Conference	Requestor	Micro	Individuals			X

(continued)

(continued)

	Publication type	Source	Task	Level	Control focus		
					Input	Behavior	Output
Gadriaju, U., Kawase, R., Dietze, S., & Demartini, G. (2015, April). Understanding malicious behavior in crowdsourcing platforms: The case of online suit'eyes. In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (pp. 1631–1640). ACM.	Conference	Requestor	Micro	Individuals	X	X	X
Gao, Y., Chen, Y., & Liu, K. R. (2015). On Cost-Effective Incentive Mechanisms in Microtask Crowdsourcing. IEEE Trans. Comput. Intellig. and AI in Games, 7(1), 3–15.	Journal	Requestor	Micro	Individuals	X		X

(continued)

(continued)

	Publication type	Source	Task	Level	Control focus		
					Input	Behavior	Output
Gould, S. J., Cox, A. L., & Brumby, D. P. (2016). Diminished control in crowdsourcing: an investigation of crowdworker multitasking behavior. <i>ACM Transactions on Computer-Human Interaction (TOCHI)</i> , 23(3), 19.	Journal	Requestor	Micro	Individuals		X	
Haas, D., Ansel, J., Gu, L., & Marcus, A. (2015). Argonaut: macrotask crowdsourcing for complex data processing. <i>Proceedings of the VLDB Endowment</i> , 8(12), 1642–1653.	Conference	Requestor	Macro	Individuals within group	X	X	X

(continued)

(continued)

	Publication type	Source	Task	Level	Control focus		
					Input	Behavior	Output
Han, S., Dai, P., Paritosh, P., & Huynh, D. (2016). Crowdsourcing Human Annotation on Web Page Structure: Infrastructure Design and Behavior-Based Quality Control. <i>ACM Transactions on Intelligent Systems and Technology (TIST)</i> , 7(4), 56.	Journal	Requestor	Micro	Individuals	X		X
Hansen, D. L., Schone, P. J., Corey, D., Reid, M., & Gehring, J. (2013, February). Quality control mechanisms for crowdsourcing: peer review, arbitration, & expertise at family search indexing. In <i>Proceedings of the 2013 conference on Computer supported cooperative work</i> (pp. 649–660). ACM.	Conference	Requestor and peer	Micro	Individuals			X

(continued)

(continued)

	Publication type	Source	Task	Level	Control focus		
					Input	Behavior	Output
Hara, K., Le, V., & Froehlich, J. (2013, April). Combining crowdsourcing and Google street view to identify street-level accessibility problems. In Proceedings of the SIGCHI conference on human factors in computing systems (pp. 631–640). ACM.	Conference	Requestor	Micro	Individuals			X
Hirth, M., Hoßfeld, T., & Tran-Gia, P. (2010). Cheat-detection mechanisms for crowdsourcing. Research Report Series, Report No 474.	Research report	Requestor	Micro	Individuals			X

(continued)

(continued)

	Publication type	Source	Task	Level	Control focus	
					Input	Output
Hirth, M., Hofßfeld, T., & Tran-Gia, P. (2011, June). Cost-optimal validation mechanisms and cheat-detection for crowdsourcing platforms. In <i>Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS)</i> , 2011 Fifth International Conference on (pp. 316–321). IEEE.	Conference	Requestor	Micro	Individuals		X
Hirth, M., Hofßfeld, T., & Tran-Gia, P. (2013). Analyzing costs and accuracy of validation mechanisms for crowdsourcing platforms. <i>Mathematical and Computer Modelling</i> , 57(11–12), 2918–2932.	Journal	Requestor	Micro	Individuals		X
Hofßfeld, T., & Keimel, C. (2014). Crowdsourcing in QoE evaluation. In <i>Quality of Experience of Experience</i> (pp. 315–327). Springer, Cham.	Book chapter	Requestor	Micro	Individuals		X

(continued)

(continued)

	Publication type	Source	Task	Level	Control focus		
					Input	Behavior	Output
Hosio, S., Goncalves, J., Lehdonvirta, V., Ferreira, D., & Kostakos, V. (2014, October). Situated crowdsourcing using a market model. In Proceedings of the 27th annual ACM symposium on User interface software and technology (pp. 55–64). ACM.	Conference	Requestor and peer	Micro	Individuals			X
Hutton, A., Liu, A., & Martin, C. E. (2012, March). Crowdsourcing Evaluations of Classifier Interpretability. In AAAI Spring Symposium: Wisdom of the Crowd.	Conference	Requestor	Micro	Individuals	X		X
Jo, J., Stevens, A., & Tan, C. (2013). A quality control model for trustworthy crowdsourcing in collaborative learning. In robot intelligence technology and applications 2012 (pp. 85–90). Springer, Berlin, Heidelberg.	Book chapter	Requestor	Micro	Individuals			X

(continued)

(continued)

	Publication type	Source	Task	Level	Control focus		
					Input	Behavior	Output
Kajino, H., Arai, H., & Kashima, H. (2014). Preserving worker privacy in crowdsourcing. <i>Data Mining and Knowledge Discovery</i> , 28(5-6), 1314-1335.	Journal	Requestor	Micro	Individuals	X		X
Kamar, E. (2016, July). Directions in Hybrid Intelligence: Complementing AI Systems with Human Intelligence. In <i>IJCAI</i> (pp. 4070-4073).	Journal	Requestor	Micro	Individuals			X
Kamar, E., Kapoor, A., Hornitz, E., & Redmond, W. A. (2013, August). Lifelong learning for acquiring the wisdom of the crowd. In <i>IJCAI</i> (Vol. 13, pp. 2313-2320).	Journal	Requestor	Micro	Individuals		X	
Kannagara, S. N., & Ugucioni, P. (2013). Risk management in crowdsourcing-based business ecosystems. <i>Technology Innovation Management Review</i> , 3(12).	Journal	Requestor	Micro	Individuals			

(continued)

(continued)

	Publication type	Source	Task	Level	Control focus		
					Input	Behavior	Output
Kazai, G. (2011, April). In search of quality in crowdsourcing for search engine evaluation. In European Conference on Information Retrieval (pp. 165–176). Springer, Berlin, Heidelberg.	Conference	Platform and requestor	Micro	Individuals	X		
Kazai, G., Kamps, J., Koolen, M., & Milic-Frayling, N. (2011, July). Crowdsourcing for book search evaluation: impact of hit design on comparative system ranking. In Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval (pp. 205–214). ACM.	Conference	Requestor	Micro	Individuals			X

(continued)

(continued)

	Publication type	Source	Task	Level	Control focus		
					Input	Behavior	Output
<p>Kazai, G., Kamps, J., & Milic-Frayling, N. (2012, October). The face of quality in crowdsourcing relevance labels: Demographics, personality and labeling accuracy. In Proceedings of the 21st ACM international conference on Information and knowledge management (pp. 2583–2586). ACM.</p>	Conference	Platform and requestor	Micro/Macro	Individuals	X		
<p>Kazai, G., Koolen, M., Kamps, J., Doucet, A., & Landoni, M. (2010, December). Overview of the INEX 2010 book track: Scaling up the evaluation using crowdsourcing. In International Workshop of the Initiative for the Evaluation of XML Retrieval (pp. 98–117). Springer, Berlin, Heidelberg.</p>	Workshop	Requestor	Micro	Individuals			

(continued)

	Publication type	Source	Task	Level	Control focus		
					Input	Behavior	Output
(continued)	Conference	Requestor	Micro	Individuals			
Kazai, G., & Zitouni, I. (2016, February). Quality management in crowdsourcing using gold judges behavior. In Proceedings of the Ninth ACM International Conference on Web Search and Data Mining (pp. 267–276). ACM.	Conference	Requestor	Micro	Individuals			
Khapra, M. M., Ramanathan, A., Kunchukuttan, A., Visweswariah, K., & Bhattacharyya, P. (2014). When Transliteration Met Crowdsourcing: An Empirical Study of Transliteration via Crowdsourcing using Efficient, Non- redundant and Fair Quality Control. Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014) (pp. 196–202).	Conference	Requestor	Micro	Individuals			

(continued)

	Publication type	Source	Task	Level	Control focus		
					Input	Behavior	Output
(continued)	Conference	Requestor	Micro	Individuals			X
Khazankin, R., Psatier, H., Schall, D., & Dustdar, S. (2011, December). Qos-based task scheduling in crowdsourcing environments. In International Conference on Service-Oriented Computing (pp. 297-311). Springer, Berlin, Heidelberg.	Conference	Requestor	Micro	Individuals			X
Lange, R., & Lange, X. (2012, March). Quality Control in Crowdsourcing: An Objective Measurement Approach to Identifying and Correcting Rater Effects in the Social Evaluation of Products and Services. In AAAI Spring Symposium: Wisdom of the Crowd (Vol. 12, p. 06).	Conference	Requestor	Micro	Individuals			X

(continued)

(continued)

	Publication type	Source	Task	Level	Control focus	
					Input	Behavior
Lasecki, W. S., & Bigham, J. P. (2012, October). Online quality control for real-time crowd captioning. In Proceedings of the 14th international ACM SIGACCESS conference on Computers and accessibility (pp. 143–150). ACM.	Conference	Requestor	Micro	Individuals		X
Lasecki, W. S., Miller, C. D., & Bigham, J. P. (2013, April). Warping time for more effective real-time crowdsourcing. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 2033–2036). ACM.	Conference	Requestor	Micro	Individuals		X

(continued)

(continued)

	Publication type	Source	Task	Level	Control focus		
					Input	Behavior	Output
Lasecki, W. S., Murray, K. I., White, S., Miller, R. C., & Bigham, J. P. (2011, October). Real-time crowd control of existing interfaces. In Proceedings of the 24th annual ACM symposium on User interface software and technology (pp. 23–32). ACM.	Conference	Requestor	Micro	Individuals		X	
Lasecki, W. S., Teevan, J., & Kamar, E. (2014, February). Information extraction and manipulation threats in crowd-powered systems. In Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing (pp. 248–256). ACM.	Conference	Requestor	Micro	Individuals			X

(continued)

(continued)

	Publication type	Source	Task	Level	Control focus		
					Input	Behavior	Output
Le, J., Edmonds, A., Hester, V., & Biewald, L. (2010, July). Ensuring quality in crowdsourced search relevance evaluation: The effects of training question distribution. In SIGIR 2010 workshop on crowdsourcing for search evaluation (Vol. 2126).	Workshop	Requestor	Micro	Individuals			X
Lee, C. Y., & Glass, J. (2011). A transcription task for crowdsourcing with automatic quality control. In Twelfth Annual Conference of the International Speech Communication Association.	Conference	Requestor	Micro	Individuals			X

(continued)

(continued)

	Publication type	Source	Task	Level	Control focus		
					Input	Behavior	Output
Li, H., Zhao, B., & Fuxman, A. (2014, April). The wisdom of minority: Discovering and targeting the right group of workers for crowdsourcing. In Proceedings of the 23rd international conference on World wide web (pp. 165–176). ACM.	Conference	Requestor	Micro	Individuals	X		X
Li, Q., Vempaty, A., Varshney, L. R., & Varshney, P. K. (2017). Multi-object classification via crowdsourcing with a reject option. IEEE Transactions on Signal Processing, 65(4), 1068–1081.	Journal	Requestor	Micro	Individuals			X

(continued)

(continued)

	Publication type	Source	Task	Level	Control focus		
					Input	Behavior	Output
Lin, C. H., & Weld, D. (2012). In Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence (UAI'12), Nando de Freitas and Kevin Murphy (Eds). AUAI Press, Arlington, Virginia, United States, 491-500.	Conference	Requestor	Micro	Individuals			X
Liu, Q., Ihler, A. T., & Steyvers, M. (2013). Scoring workers in crowdsourcing: How many control questions are enough?. In Advances in Neural Information Processing Systems (pp. 1914-1922).	Journal	Requestor	Micro	Individuals			X

(continued)

(continued)

	Publication type	Source	Task	Level	Control focus		
					Input	Behavior	Output
Liu, Z., Shabani, S., Balet, N. G., Sokhn, M., & Cretton, F. (2018, January). How to motivate participation and improve quality of crowdsourcing when building accessibility maps. In <i>Consumer Communications & Networking Conference (CCNC), 2018 15th IEEE Annual (pp. 1–6)</i> . IEEE.	Conference	Requestor	Micro	Individuals	X		X
Loni, B., Menendez, M., Georgescu, M., Galli, L., Massari, C., Altingovde, I. S., ... & Larson, M. (2013, February). Fashion-focused creative commons social dataset. In <i>Proceedings of the 4th ACM Multimedia Systems Conference (pp. 72–77)</i> . ACM.	Conference	Requestor	Micro	Individuals			X

(continued)

(continued)

	Publication type	Source	Task	Level	Control focus		
					Input	Behavior	Output
<p>Massung, E., Coyle, D., Cater, K. F., Jay, M., & Preist, C. (2013, April). Using crowdsourcing to support pro- environmental community activism. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 371–380). ACM.</p>	Conference	Requestor	Micro	Individuals			X
<p>McGraw, I., & Polifroni, J. (2013). How to Control and Utilize Crowd-Collected Speech. In Eskenazi, M., Levow, G., Meng, H., Parent, G., Suendermann, D. (eds) Crowdsourcing for Speech Processing: Applications to Data Collection, Transcription and Assessment, 106–136.</p>	Book chapter	Requestor	Micro	Individuals			X

(continued)

	Publication type	Source	Task	Level	Control focus		
					Input	Behavior	Output
(continued)	Conference	Requestor	Micro	Individuals			X
Melchior, P., Sheldon, E., Drllica-Wagner, A., Rykoff, E. S., Abbott, T. M. C., Abdalla, F. B.,... & Rosell, A. C. (2016). Crowdsourcing quality control for Dark Energy Survey images. <i>Astronomy and Computing</i> , 16, 99–108.	Conference	Requestor	Micro	Individuals			X
Munro, R., Bethard, S., Kuperman, V., Lai, V. T., Melnick, R., Potts, C.,... & Tily, H. (2010, June). Crowdsourcing and language studies: the new generation of linguistic data. In <i>Proceedings of the NAACL HLT 2010 workshop on creating speech and language data with Amazon's Mechanical Turk</i> (pp. 122–130). Association for Computational Linguistics.	Conference	Requestor	Micro	Individuals	X		X

(continued)

(continued)

	Publication type	Source	Task	Level	Control focus	
					Input	Behavior
Negri, M., Bentivogli, L., Mehdad, Y., Giampiccolo, D., & Marchetti, A. (2011, July). Divide and conquer: Crowdsourcing the creation of cross-lingual textual entailment corpora. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (pp. 670–679). Association for Computational Linguistics.	Conference	Requestor	Micro	Individuals		X
Oleson, D., Sorokin, A., Laughlin, G. P., Hester, V., Le, J., & Biewald, L. (2011). Programmatic Gold: Targeted and Scalable Quality Assurance in Crowdsourcing. Human computation, 11(11).	Journal	Requestor	Micro	Individuals		X

(continued)

(continued)

	Publication type	Source	Task	Level	Control focus	
					Input	Behavior
Otani, N., Baba, Y., & Kashima, H. (2016). Quality control of crowdsourced classification using hierarchical class structures. <i>Expert Systems with Applications</i> , 58, 155–163.	Journal	Requestor	Micro	Individuals		X
Oyama, S., Baba, Y., Ohmukai, I., Dokoshi, H., & Kashima, H. (2015). From one star to three stars: Upgrading legacy open data using crowdsourcing. <i>IEEE International Conference on Data Science and Advanced Analytics</i> ; http://hdl.handle.net/2115/65226	Conference	Requestor	Micro	Individuals		X

(continued)

(continued)

	Publication type	Source	Task	Level	Control focus		
					Input	Behavior	Output
Oyama, S., Baba, Y., Sakurai, Y., & Kashima, H. (2013, August). Accurate Integration of Crowdsourced Labels Using Workers' Self-reported Confidence Scores. Twenty-Third International Joint Conference on Artificial Intelligence (pp. 2554–2560).	Journal	Requestor	Micro	Individuals			X
Paul, S. A., Hong, L., & Chi, E. H. (2011). What is a question? Crowdsourcing tweet categorization. HCOMP Workshop CHI 2011.	Workshop	Requestor	Micro	Individuals	X		X

(continued)

	Publication type	Source	Task	Level	Control focus		
					Input	Behavior	Output
(continued)	Workshop	Requestor	Micro	Individuals			
Post, M., Callison-Burch, C., & Osborne, M. (2012, June). Constructing parallel corpora for six indian languages via crowdsourcing. In Proceedings of the Seventh Workshop on Statistical Machine Translation (pp. 401–409). Association for Computational Linguistics.							
Qiu, C., Squicciarini, A. C., Carminati, B., Caverlee, J., & Khare, D. R. (2016, October). Crowdselect: Increasing accuracy of crowdsourcing tasks through behavior prediction and user selection. In Proceedings of the 25th ACM International Conference on Information and Knowledge Management (pp. 539–548). ACM.	Conference	Requestor	Micro	Individuals	X		X

(continued)

(continued)

	Publication type	Source	Task	Level	Control focus		
					Input	Behavior	Output
Rhyn, M., & Blohm, I. (2017) A Machine Learning Approach for Classifying Textual Data in Crowdsourcing, in Leimeister, J.M.; Brenner, W. (Hrsg.): Proceedings der 13. Internationalen Tagung Wirtschaftsinformatik (WI 2017), St. Gallen, S. 1171–1185.	Conference	Requestor	Micro	Individuals			X
Riccardi, G., Ghosh, A., Chowdhury, S. A., & Bayer, A. O. (2013, August). Motivational feedback in crowdsourcing: a case study in speech transcription. In INTERSPEECH (pp. 1111–1115).	Conference	Requestor	Micro	Individuals			X

(continued)

(continued)

	Publication type	Source	Task	Level	Control focus		
					Input	Behavior	Output
Riegler, M., Gaddam, V. R., Larson, M., Eg, R., Halvorsen, P., & Griwodz, C. (2016, June). Crowdsourcing as self-fulfilling prophecy: Influence of discarding workers in subjective assessment tasks. In Content-Based Multimedia Indexing (CBMI), 2016 14th International Workshop on (pp. 1–6). IEEE.	Workshop	Requestor	Micro	Individuals			
Salk, C. F., Sturm, T., See, L., Fritz, S., & Perget, C. (2016). Assessing quality of volunteer crowdsourcing contributions: lessons from the Cropland Capture game. International Journal of Digital Earth, 9(4), 410–426.	Journal	Requestor	Micro	Individuals			X

(continued)

	Publication type	Source	Task	Level	Control focus		
					Input	Behavior	Output
(continued)							
Satzger, B., Psailer, H., Schall, D., & Dustdar, S. (2013). Auction-based crowdsourcing supporting skill management. <i>Information Systems</i> , 38(4), 547–560.	Journal	Platform and requestor	Micro	Individuals	X		X
See, L., Schepaschenko, D., Lesiv, M., McCallum, I., Fritz, S., Comber, A. & Siraj, M. A. (2015). Building a hybrid land cover map with crowdsourcing and geographically weighted regression. <i>ISPRS Journal of Photogrammetry and Remote Sensing</i> , 103, 48–56.	Journal	Requestor	Micro	Individuals			X
Schmitz, H., & Lykourantzou, I. (2018). Online Sequencing of Non-Decomposable Macrotasks in Expert Crowdsourcing. <i>ACM Transactions on Social Computing</i> , 1(1), 1.	Journal	Requestor	Macro	Individuals within group		X	

(continued)

(continued)

	Publication type	Source	Task	Level	Control focus		
					Input	Behavior	Output
Sorokin, A., Berenson, D., Srinivasa, S. S., & Hebert, M. (2010, October). People helping robots helping people: Crowdsourcing for grasping novel objects. In 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems (pp. 2117–2122). IEEE.	Conference	Requestor	Micro	Individuals			X
Sprugnoli, R., Moretti, G., Fuoli, M., Giuliani, D., Bentivogli, L., Pianta, E.,... & Brugnara, F. (2013, May). Comparing two methods for crowdsourcing speech transcription. In 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (pp. 8116–8120). IEEE.	Conference	Requestor	Micro	Individuals			X

(continued)

(continued)

	Publication type	Source	Task	Level	Control focus		
					Input	Behavior	Output
Stolee, K. T., & Elbaum, S. (2010, September). Exploring the use of crowdsourcing to support empirical studies in software engineering. In Proceedings of the 2010 ACM-IEEE international symposium on Empirical software engineering and measurement (p. 35). ACM.	Conference	Platform and requestor	Micro	Individuals	X		
Su, H., Deng, J., & Fei-Fei, L. (2012, July). Crowdsourcing annotations for visual object detection. In Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence (Vol. 1, No. 2).	Conference	Requestor	Micro	Individuals	X		X

(continued)

(continued)

	Publication type	Source	Task	Level	Control focus		
					Input	Behavior	Output
Tai, L., Chuang, Z., Tao, X., Ming, W., & Jingjing, X. (2011). Quality control of crowdsourcing through workers experience. In Proceedings of the ACM SIGIR workshop on crowdsourcing for information retrieval.	Conference	Requestor	Micro	Individuals	X		X
Tang, W., & Lease, M. (2011, July). Semi-supervised consensus labeling for crowdsourcing. In SIGIR 2011 workshop on crowdsourcing for information retrieval (CIR) (pp. 1–6).	Workshop	Requestor	Micro	Individuals			X

(continued)

(continued)

	Publication type	Source	Task	Level	Control focus	
					Input	Behavior
Tran-Thanh, L., Huynh, T., D., Rosenfeld, A., Ramechurn, S. D., & Jennings, N. R. (2014, May). BudgetFix: budget limited crowdsourcing for interdependent task allocation with quality guarantees. In Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems (pp. 477–484). International Foundation for Autonomous Agents and Multiagent Systems.	Conference	Requestor	Micro	Individuals		X
Trompette, P., Chanal, V., & Pelissier, C. (2008, July). Crowdsourcing as a way to access external knowledge for innovation. In 24 th EGOS Colloquium.	Conference	Requestor	Macro	Individuals		

(continued)

	Publication type	Source	Task	Level	Control focus		
					Input	Behavior	Output
(continued)							
UI Hassan, U., Zaveri, A., Marx, E., Curry, E., & Lehmann, J. (2016, November). ACRyLIQ: Leveraging DBpedia for adaptive crowdsourcing in linked data quality assessment. In European Knowledge Acquisition Workshop (pp. 681-696). Springer, Cham.	Workshop	Requestor	Micro	Individuals			X
Vempaty, A., Varshney, L. R., & Varshney, P. K. (2014). Reliable crowdsourcing for multi-class labeling using coding theory. IEEE Journal of Selected Topics in Signal Processing, 8(4), 667-679.	Journal	Requestor	Micro	Individuals			X
Venetis, P., & Garcia-Molina, H. (2012, August). Quality control for comparison microtasks. In Proceedings of the first international workshop on crowdsourcing and data mining (pp. 15-21). ACM.	Conference	Requestor	Micro	Individuals			X

(continued)

(continued)

	Publication type	Source	Task	Level	Control focus		
					Input	Behavior	Output
Vliegandhart, R., Larson, M., Kofler, C., Eickhoff, C., & Pouwelse, J. (2011, February). Investigating factors influencing crowdsourcing tasks with high imaginative load. In Proceedings of the Workshop on Crowdsourcing for Search and Data Mining (CSDM) at the Fourth ACM International Conference on Web Search and Data Mining (pp. 2730).	Conference	Requestor	Micro	Individuals	X		X
Wais, P., Lingamneni, S., Cook, D., Fennell, J., Goldenberg, B., Lubarov, D., ... & Simons, H. (2010). Towards building a high-quality workforce with mechanical turk. Proceedings of computational social science and the wisdom of crowds (NIPS), 1–5.	Conference	Requestor	Micro	Individuals	X		X

(continued)

(continued)

	Publication type	Source	Task	Level	Control focus		
					Input	Behavior	Output
Wang, S., Huang, C. R., Yao, Y., & Chan, A. (2014). Exploring mental lexicon in an efficient and economic way: Crowdsourcing method for linguistic experiments. In Proceedings of the 4th Workshop on Cognitive Aspects of the Lexicon (CogALex) (pp. 105–113).	Conference	Requestor	Micro	Individuals	X		X
Wu, C. C., Chen, K. T., Chang, Y. C., & Lei, C. L. (2013). Crowdsourcing multimedia QoE evaluation: A trusted framework. IEEE transactions on multimedia, 15(5), 1121–1137.	Journal	Requestor	Micro	Individuals			X

(continued)

(continued)

	Publication type	Source	Task	Level	Control focus		
					Input	Behavior	Output
Xia, T., Zhang, C., Xie, J., & Li, T. (2012, September). Real-time quality control for crowdsourcing relevance evaluation. In Network Infrastructure and Digital Content (IC-NIDC), 2012 3rd IEEE International Conference on (pp. 535–539). IEEE.	Conference	Requestor	Micro	Individuals	X	X	X
Yung, D., Li, M. L., & Chang, S. (2014). Evolutionary approach for crowdsourcing quality control. Journal of Visual Languages & Computing, 25(6), 879–890.	Journal	Requestor	Micro	Individuals		X	X

(continued)

	Publication type	Source	Task	Level	Control focus		
					Input	Behavior	Output
(continued)	Conference	Requestor	Micro	Individuals			X
Zaidan, O. F., & Callison-Burch, C. (2011, June). Crowdsourcing translation: Professional quality from non-professionals. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1 (pp. 1220–1229). Association for Computational Linguistics.							
Zhai, H., Lingren, T., Deleger, L., Li, Q., Kaiser, M., Stoutenborough, L., & Solti, I. (2013). Web high-quality gold standard development in clinical natural language processing. Journal of medical Internet research, 15(4).	Journal	Platform and requestor	Micro	Individuals	X		X

(continued)

(continued)

	Publication type	Source	Task	Level	Control focus		
					Input	Behavior	Output
Zhang, G., & Chen, H. (2013, October). Quality control for crowdsourcing with spatial and temporal distribution. In International Conference on Internet and Distributed Computing Systems (pp. 169–182). Springer, Berlin, Heidelberg.	Conference	Requestor	Micro	Individuals			
Zhang, G., & Chen, H. (2013, December). Quality control of massive data for crowdsourcing in location-based services. In International Conference on Algorithms and Architectures for Parallel Processing (pp. 112–121). Springer, Cham.	Conference	Requestor	Micro	Individuals			X
Zogaj, S., & Bretschneider, U. (2014). Analyzing governance mechanisms for crowdsourcing information systems: a multiple case analysis. ECIS 2014	Conference	Requestor	Micro	Individuals			X

References

- Aker, A., El-Haj, M., Albakour, M. D., & Kruschwitz, U. (2012). Assessing crowdsourcing quality through objective tasks. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation* (pp. 1456–1461).
- Ashikawa, M., Kawamura, T., & Ohsuga, A. (2015). Deployment of private crowdsourcing system with quality control methods. In *2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)* (Vol. 1, pp. 9–16). IEEE.
- Baba, Y., & Kashima, H. (2013, August). Statistical quality estimation for general crowdsourcing tasks. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 554–562). ACM.
- Baba, Y., Kashima, H., Kinoshita, K., Yamaguchi, G., & Akiyoshi, Y. (2013, June). Leveraging crowdsourcing to detect improper tasks in crowdsourcing marketplaces. In *Twenty-fifth Innovative Applications of Artificial Intelligence Conference* (pp. 1487–1492).
- Baba, Y., Kashima, H., Kinoshita, K., Yamaguchi, G., & Akiyoshi, Y. (2014). Leveraging non-expert crowdsourcing workers for improper task detection in crowdsourcing marketplaces. *Expert Systems with Applications*, *41*(6), 2678–2687.
- Bell, S., & Bala, K. (2015). Learning visual similarity for product design with convolutional neural networks. *ACM Transactions on Graphics (TOG)*, *34*(4), 98.
- Bontcheva, K., Roberts, I., Derczynski, L., & Rout, D. (2014). The GATE crowdsourcing plugin: Crowdsourcing annotated corpora made easy. In *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 97–100).
- Bozzon, A., Brambilla, M., Ceri, S., & Mauri, A. (2013, May). Reactive crowdsourcing. In *Proceedings of the 22nd International Conference on World Wide Web* (pp. 153–164). ACM.
- Bozzon, A., Brambilla, M., Ceri, S., Mauri, A., & Volonterio, R. (2014, July). Pattern-based specification of crowdsourcing applications. In *International Conference on Web Engineering* (pp. 218–235). Cham: Springer.
- Bragg, J., & Weld, D. S. (2013, November). Crowdsourcing multi-label classification for taxonomy creation. In *First AAAI Conference on Human Computation and Crowdsourcing*.
- Cardinal, L. B., Kreutzer, M., & Miller, C. C. (2017). An aspirational view of organizational control research: Re-invigorating empirical work to better meet the challenges of 21st century organizations. *Academy of Management Annals*, *11*(2), 559–592.
- Cardinal, L. B., Sitkin, S. B., & Long, C. P. (2004). Balancing and rebalancing in the creation and evolution of organizational control. *Organization Science*, *15*, 411–431.
- Cardinal, L. B., Sitkin, S. B., & Long, C. P. (2010). A configurational theory of control. In S. B. Sitkin, L. B. Cardinal, & K. M. Bijlsma-Frankema (Eds.), *Organizational control* (pp. 51–79). Cambridge, UK: Cambridge University Press.
- Carpenter, M. A., Bauer, T., Erdogan, B., & Short, J. (2010). *Principles of management*. Flatworld Knowledge.
- Causser, T., Tonra, J., & Wallace, V. (2012). Transcription maximized; expense minimized? Crowdsourcing and editing the collected works of Jeremy Bentham. *Literary and Linguistic Computing*, *27*(2), 119–137.
- Chang, D., Chen, C. H., & Lee, K. M. (2014). A crowdsourcing development approach based on a neuro-fuzzy network for creating innovative product concepts. *Neurocomputing*, *142*, 60–72.
- Chatman, J. A. (2010). Norms in mixed sex and mixed race work groups. *Academy of Management Annals*, *4*(1), 447–484.
- Chen, Z., Fu, R., Zhao, Z., Liu, Z., Xia, L., Chen, L., et al. (2014). gMission: A general spatial crowdsourcing platform. *Proceedings of the VLDB Endowment*, *7*(13), 1629–1632.
- Cheng, J., Teevan, J., Iqbal, S. T., & Bernstein, M. S. (2015, April). Break it down: A comparison of macro-and microtasks. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (pp. 4061–4064). ACM.

- Chiu, C. M., Liang, T. P., & Turban, E. (2014). What can crowdsourcing do for decision support? *Decision Support Systems*, 65, 40–49.
- Choudhury, V., & Sabherwal, R. (2003). Portfolios of control in outsourced software development projects. *Information Systems Research*, 14(3), 291–314.
- Chung, M. J. Y., Forbes, M., Cakmak, M., & Rao, R. P. (2014, May). Accelerating imitation learning through crowdsourcing. In *ICRA* (pp. 4777–4784).
- Dai, P., Lin, C. H., & Weld, D. S. (2013). POMDP-based control of workflows for crowdsourcing. *Artificial Intelligence*, 202, 52–85.
- Dai, P., Rzeszotarski, J. M., Paritosh, P., & Chi, E. H. (2015, February). And now for something completely different: Improving crowdsourcing workflows with micro-diversions. In *Proceedings of the 18th ACM Conference on Computer-Supported Cooperative Work & Social Computing* (pp. 628–638). ACM.
- Daniel, F., Kucherbaev, P., Cappiello, C., Benatallah, B., & Allahbakhsh, M. (2018). Quality control in crowdsourcing: A survey of quality attributes, assessment techniques, and assurance actions. *ACM Computing Surveys (CSUR)*, 51(1), 7.
- de Herrera, A. G. S., Foncubierta-Rodríguez, A., Markonis, D., Schaer, R., & Müller, H. (2014, September). Crowdsourcing for medical image classification. In *Annual Congress SGMI* (Vol. 2014).
- Deng, J., Krause, J., & Fei-Fei, L. (2013). Fine-grained crowdsourcing for fine-grained recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 580–587).
- Dennis, A. R., Robert, L. P., Kowalczyk, S. T., Curtis, A., & Hasty, B. K. (2012). Trust is in the eye of the beholder: A vignette study of postevent behavioral controls' effects on individual trust in virtual teams. *Information Systems Research*, 23(2), 546–558.
- Difallah, D. E., Demartini, G., & Cudré-Mauroux, P. (2012, April). Mechanical cheat: Spamming schemes and adversarial techniques on crowdsourcing platforms. In *CrowdSearch 2010 Workshop at WWW 2012* (pp. 26–30).
- Duan, L., Oyama, S., Sato, H., & Kurihara, M. (2014). Separate or joint? Estimation of multiple labels from crowdsourced annotations. *Expert Systems with Applications*, 41(13), 5723–5732.
- Eickhoff, C., & de Vries, A. (2011, February). How crowdsourcable is your task? In *Proceedings of the Workshop on Crowdsourcing for Search and Data Mining (CSDM) at the Fourth ACM International Conference on Web Search and Data Mining (WSDM)* (pp. 11–14).
- Eickhoff, C., & de Vries, A. P. (2013). Increasing cheat robustness of crowdsourcing tasks. *Information Retrieval*, 16(2), 121–137.
- Eisenhardt, K. M. (1985). Control: Organizational and economic approaches. *Management Science*, 31, 134–149.
- Fan, J., Li, G., Ooi, B. C., Tan, K. L., & Feng, J. (2015, May). iCrowd: An adaptive crowdsourcing framework. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data* (pp. 1015–1030). ACM.
- Fang, Y., Sun, H., Li, G., Zhang, R., & Huai, J. (2016, April). Effective result inference for context-sensitive tasks in crowdsourcing. In *International Conference on Database Systems for Advanced Applications* (pp. 33–48). Cham: Springer.
- Filatova, E. (2012, May). Irony and sarcasm: Corpus generation and analysis using crowdsourcing. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation* (pp. 392–398).
- Finin, T., Murnane, W., Karandikar, A., Keller, N., Martineau, J., & Dredze, M. (2010, June). Annotating named entities in Twitter data with crowdsourcing. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk* (pp. 80–88). Association for Computational Linguistics.
- Foncubierta Rodríguez, A., & Müller, H. (2012, October). Ground truth generation in medical imaging: A crowdsourcing-based iterative approach. In *Proceedings of the ACM Multimedia 2012 Workshop on Crowdsourcing for Multimedia* (pp. 9–14). ACM.

- Franklin, M. J., Kossmann, D., Kraska, T., Ramesh, S., & Xin, R. (2011, June). CrowdDB: Answering queries with crowdsourcing. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data* (pp. 61–72). ACM.
- Fu, W. T., & Liao, V. (2011, March). Crowdsourcing quality control of online information: A quality-based cascade model. In *International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction* (pp. 147–154). Berlin, Heidelberg: Springer.
- Gasiraju, U., Kawase, R., Dietze, S., & Demartini, G. (2015). Understanding malicious behavior in crowdsourcing platforms: The case of online surveys. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (pp. 1631–1640). ACM.
- Gao, Y., Chen, Y., & Liu, K. R. (2015). On cost-effective incentive mechanisms in microtask crowdsourcing. *IEEE Transactions in Computational Intelligence and AI in Games*, 7(1), 3–15.
- Gould, S. J., Cox, A. L., & Brumby, D. P. (2016). Diminished control in crowdsourcing: An investigation of crowdworker multitasking behavior. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 23(3), 19.
- Haas, D., Ansel, J., Gu, L., & Marcus, A. (2015). Argonaut: Macrotask crowdsourcing for complex data processing. *Proceedings of the VLDB Endowment*, 8(12), 1642–1653.
- Han, S., Dai, P., Paritosh, P., & Huynh, D. (2016). Crowdsourcing human annotation on web page structure: Infrastructure design and behavior-based quality control. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 7(4), 56.
- Hansen, D. L., Schone, P. J., Corey, D., Reid, M., & Gehring, J. (2013, February). Quality control mechanisms for crowdsourcing: Peer review, arbitration, & expertise at family search indexing. In *Proceedings of the 2013 Conference on Computer-Supported Cooperative Work* (pp. 649–660). ACM.
- Hara, K., Le, V., & Froehlich, J. (2013, April). Combining crowdsourcing and Google Street View to identify street-level accessibility problems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 631–640). ACM.
- Hirth, M., Hoßfeld, T., & Tran-Gia, P. (2010). Cheat-detection mechanisms for crowdsourcing. Research report series, report No. 474.
- Hirth, M., Hoßfeld, T., & Tran-Gia, P. (2011, June). Cost-optimal validation mechanisms and cheat-detection for crowdsourcing platforms. In *2011 Fifth International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS)* (pp. 316–321). IEEE.
- Hirth, M., Hoßfeld, T., & Tran-Gia, P. (2013). Analyzing costs and accuracy of validation mechanisms for crowdsourcing platforms. *Mathematical and Computer Modelling*, 57(11–12), 2918–2932.
- Hoßfeld, T., & Keimel, C. (2014). Crowdsourcing in QoE evaluation. In *Quality of experience* (pp. 315–327). Cham: Springer.
- Homan, A. C., van Knippenberg, D., Van Kleef, G. A., & De Dreu, C. K. W. (2007). Bridging faultlines by valuing diversity: The effects of diversity beliefs on information elaboration and performance in diverse work groups. *Journal of Applied Psychology*, 92, 1189–1199.
- Hosio, S., Goncalves, J., Lehdonvirta, V., Ferreira, D., & Kostakos, V. (2014, October). Situated crowdsourcing using a market model. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology* (pp. 55–64). ACM.
- Hutton, A., Liu, A., & Martin, C. E. (2012, March). Crowdsourcing evaluations of classifier interpretability. In *AAAI Spring Symposium: Wisdom of the Crowd*.
- Jaworski, B. J., & Kohli, A. K. (1993). Market orientation: Antecedents and consequences. *Journal of Marketing*, 57, 53–70.
- Jo, J., Stevens, A., & Tan, C. (2013). A quality control model for trustworthy crowdsourcing in collaborative learning. In *Robot intelligence technology and applications 2012* (pp. 85–90). Berlin, Heidelberg: Springer.
- Kajino, H., Arai, H., & Kashima, H. (2014). Preserving worker privacy in crowdsourcing. *Data Mining and Knowledge Discovery*, 28(5–6), 1314–1335.
- Kamar, E. (2016, July). Directions in hybrid intelligence: Complementing AI systems with human intelligence. In *IJCAI* (pp. 4070–4073).

- Kamar, E., Kapoor, A., Horvitz, E., & Redmond, W. A. (2013, August). Lifelong learning for acquiring the wisdom of the crowd. In *IJCAI* (Vol. 13, pp. 2313–2320).
- Kannangara, S. N., & Uguccioni, P. (2013). Risk management in crowdsourcing-based business ecosystems. *Technology Innovation Management Review*, 3(12).
- Kazai, G. (2011, April). In search of quality in crowdsourcing for search engine evaluation. In *European Conference on Information Retrieval* (pp. 165–176). Berlin, Heidelberg: Springer.
- Kazai, G., Kamps, J., Koolen, M., & Milic-Frayling, N. (2011, July). Crowdsourcing for book search evaluation: Impact of hit design on comparative system ranking. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 205–214). ACM.
- Kazai, G., Kamps, J., & Milic-Frayling, N. (2012, October). The face of quality in crowdsourcing relevance labels: Demographics, personality and labeling accuracy. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management* (pp. 2583–2586). ACM.
- Kazai, G., Koolen, M., Kamps, J., Doucet, A., & Landoni, M. (2010, December). Overview of the INEX 2010 book track: Scaling up the evaluation using crowdsourcing. In *International Workshop of the Initiative for the Evaluation of XML Retrieval* (pp. 98–117). Berlin, Heidelberg: Springer.
- Kazai, G., & Zitouni, I. (2016, February). Quality management in crowdsourcing using gold judges behavior. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining* (pp. 267–276). ACM.
- Khapra, M. M., Ramanathan, A., Kunchukuttan, A., Visweswariah, K., & Bhattacharyya, P. (2014). When transliteration met crowdsourcing: An empirical study of transliteration via crowdsourcing using efficient, non-redundant and fair quality control. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)* (pp. 196–202).
- Khazankin, R., Psai, H., Schall, D., & Dustdar, S. (2011, December). Qos-based task scheduling in crowdsourcing environments. In *International Conference on Service-oriented Computing* (pp. 297–311). Berlin, Heidelberg: Springer.
- Kim, S., Marquis, E., Alahmad, R., Pierce, C., & Robert, L. P. (2018). The impacts of platform quality on gig workers' autonomy and satisfaction. In *Proceedings of the 21th ACM Conference on Computer-supported Cooperative Work and Social Computing Companion*. Jersey City, NJ, USA.
- Kirsch, L. J. (1997). Portfolios of control modes and IS project management. *Information Systems Research*, 8(3), 215–239.
- Kirsch, L. J., Ko, D. G., & Haney, M. H. (2010). Investigating the antecedents of team-based clan control: Adding social capital as a predictor. *Organization Science*, 21(2), 469–489.
- Lange, R., & Lange, X. (2012, March). Quality control in crowdsourcing: An objective measurement approach to identifying and correcting rater effects in the social evaluation of products and services. In *AAAI Spring Symposium: Wisdom of the Crowd* (Vol. 12, p. 6).
- Lasecki, W. S., & Bigham, J. P. (2012, October). Online quality control for real-time crowd captioning. In *Proceedings of the 14th international ACM SIGACCESS Conference on Computers and Accessibility* (pp. 143–150). ACM.
- Lasecki, W. S., Miller, C. D., & Bigham, J. P. (2013, April). Warping time for more effective real-time crowdsourcing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 2033–2036). ACM.
- Lasecki, W. S., Murray, K. I., White, S., Miller, R. C., & Bigham, J. P. (2011, October). Real-time crowd control of existing interfaces. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology* (pp. 23–32). ACM.
- Lasecki, W. S., Teevan, J., & Kamar, E. (2014, February). Information extraction and manipulation threats in crowd-powered systems. In *Proceedings of the 17th ACM Conference on Computer-supported Cooperative Work & Social Computing* (pp. 248–256). ACM.

- Le, J., Edmonds, A., Hester, V., & Biewald, L. (2010, July). Ensuring quality in crowdsourced search relevance evaluation: The effects of training question distribution. In *SIGIR 2010 Workshop on Crowdsourcing for Search Evaluation* (Vol. 2126).
- Lee, C. Y., & Glass, J. (2011). *A transcription task for crowdsourcing with automatic quality control*. Paper Presented at the Twelfth Annual Conference of the International Speech Communication Association.
- Li, H., Zhao, B., & Fuxman, A. (2014, April). The wisdom of minority: Discovering and targeting the right group of workers for crowdsourcing. In *Proceedings of the 23rd International Conference on World Wide Web* (pp. 165–176). ACM.
- Li, Q., Vempaty, A., Varshney, L. R., & Varshney, P. K. (2017). Multi-object classification via crowdsourcing with a reject option. *IEEE Transactions on Signal Processing*, 65(4), 1068–1081.
- Lin, C. H., & Weld, D. (2012). In N. de Freitas & K. Murphy (Eds.), *Proceedings of the Twenty-eighth Conference on Uncertainty in Artificial Intelligence (UAI'12)* (pp. 491–500). Arlington, VA: AUA Press.
- Liu, Q., Ihler, A. T., & Steyvers, M. (2013). Scoring workers in crowdsourcing: How many control questions are enough? In *Advances in neural information processing systems* (pp. 1914–1922).
- Liu, S. (2015). Effects of control on the performance of information systems projects: The moderating role of complexity risk. *Journal of Operations Management*, 36, 46–62.
- Liu, Z., Shabani, S., Balet, N. G., Sokhn, M., & Cretton, F. (2018, January). How to motivate participation and improve quality of crowdsourcing when building accessibility maps. In *2018 15th IEEE Annual Consumer Communications & Networking Conference (CCNC)* (pp. 1–6). IEEE.
- Loni, B., Menendez, M., Georgescu, M., Galli, L., Massari, C., Altingovde, I. S., ... & Larson, M. (2013, February). Fashion-focused creative commons social dataset. In *Proceedings of the 4th ACM Multimedia Systems Conference* (pp. 72–77). ACM.
- Malhotra, A., & Majchrzak, A. (2014). Managing crowds in innovation challenges. *California Management Review*, 56(4), 103–123.
- Maruping, L. M., Venkatesh, V., & Agarwal, R. (2009). A control theory perspective on agile methodology use and changing user requirements. *Information Systems Research*, 20(3), 377–399.
- Massung, E., Coyle, D., Cater, K. F., Jay, M., & Preist, C. (2013, April). Using crowdsourcing to support pro-environmental community activism. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 371–380). ACM.
- Mays, N., Roberts, E., & Popay, J. (2001). Synthesising research evidence. In N. Fulop, P. Allen, A. Clarke, & N. Black (Eds.), *Studying the organisation and delivery of health services: Research methods* (pp. 188–219). London: Routledge.
- McGraw, I., & Polifroni, J. (2013). How to control and utilize crowd-collected speech. In M. Eskenazi, G. Levow, H. Meng, G. Parent, & D. Suendermann (Eds.), *Crowdsourcing for speech processing: Applications to data collection, transcription and assessment* (pp. 106–136). Chichester, UK: Wiley.
- Melchior, P., Sheldon, E., Drlica-Wagner, A., Rykoff, E. S., Abbott, T. M. C., Abdalla, F. B., et al. (2016). Crowdsourcing quality control for Dark Energy Survey images. *Astronomy and Computing*, 16, 99–108.
- Munro, R., Bethard, S., Kuperman, V., Lai, V. T., Melnick, R., Potts, C., ... & Tily, H. (2010, June). Crowdsourcing and language studies: The new generation of linguistic data. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk* (pp. 122–130). Association for Computational Linguistics.
- Negri, M., Bentivogli, L., Mehdad, Y., Giampiccolo, D., & Marchetti, A. (2011, July). Divide and conquer: Crowdsourcing the creation of cross-lingual textual entailment corpora. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 670–679). Association for Computational Linguistics.
- Oleson, D., Sorokin, A., Laughlin, G. P., Hester, V., Le, J., & Biewald, L. (2011). Programmatic gold: Targeted and scalable quality assurance in crowdsourcing. *Human Computation*, 11(11).

- Otani, N., Baba, Y., & Kashima, H. (2016). Quality control of crowdsourced classification using hierarchical class structures. *Expert Systems with Applications*, 58, 155–163.
- Ouchi, W. G. (1979). A conceptual framework for the design of organizational control mechanisms. *Management Science*, 25(9), 833–848.
- Ouchi, W. G. (1980). Markets, bureaucracies, and clans. *Administrative Science Quarterly*, 25(1), 129–141.
- Ouchi, W. G., & Price, R. L. (1978). Hierarchies, clans, and theory Z: A new perspective on organization development. *Organizational Dynamics*, 7(2), 25–44.
- Oyama, S., Baba, Y., Ohmukai, I., Dokoshi, H., & Kashima, H. (2015). From one star to three stars: Upgrading legacy open data using crowdsourcing. In *IEEE International Conference on Data Science and Advanced Analytics* (pp. 1–9). IEEE.
- Oyama, S., Baba, Y., Sakurai, Y., & Kashima, H. (2013, August). Accurate integration of crowdsourced labels using workers' self-reported confidence scores. In *Twenty-third International Joint Conference on Artificial Intelligence* (pp. 2554–2560).
- Paul, S. A., Hong, L., & Chi, E. H. (2011). *What is a question? Crowdsourcing tweet categorization*. Paper Presented at HCOMP Workshop CHI 2011.
- Peterson, J., Pearce, P. F., Ferguson, L. A., & Langford, C. A. (2017). Understanding scoping reviews: Definition, purpose, and process. *Journal of the American Association of Nurse Practitioners*, 29(1), 12–16.
- Piccoli, G., & Ives, B. (2003). Trust and the unintended effects of behavior control in virtual teams. *MIS Quarterly*, 27(3), 365–395.
- Post, M., Callison-Burch, C., & Osborne, M. (2012, June). Constructing parallel corpora for six Indian languages via crowdsourcing. In *Proceedings of the Seventh Workshop on Statistical Machine Translation* (pp. 401–409). Association for Computational Linguistics.
- Qiu, C., Squicciarini, A. C., Carminati, B., Caverlee, J., & Khare, D. R. (2016, October). Crowdselect: Increasing accuracy of crowdsourcing tasks through behavior prediction and user selection. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management* (pp. 539–548). ACM.
- Rhyn, M., & Blohm, I. (2017). A machine learning approach for classifying textual data in crowdsourcing. In J. M. Leimeister & W. Brenner, W. (Eds.), *Proceedings der 13. Internationalen Tagung Wirtschaftsinformatik (WI 2017)* (pp. 1171–1185).
- Riccardi, G., Ghosh, A., Chowdhury, S. A., & Bayer, A. O. (2013, August). Motivational feedback in crowdsourcing: A case study in speech transcription. In *INTERSPEECH* (pp. 1111–1115).
- Riegler, M., Gaddam, V. R., Larson, M., Eg, R., Halvorsen, P., & Griwodz, C. (2016, June). Crowdsourcing as self-fulfilling prophecy: Influence of discarding workers in subjective assessment tasks. In *2016 14th International Workshop on Content-Based Multimedia Indexing (CBMI)* (pp. 1–6). IEEE.
- Robert, L. P. (2016). Monitoring and trust in virtual teams. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW 2016)*. ACM.
- Robert, L. P., Jr., Dennis, A. R., & Ahuja, M. K. (2008). Social capital and knowledge integration in digitally enabled teams. *Information Systems Research*, 19(3), 314–334.
- Salehi, N., McCabe, A., Valentine, M., & Bernstein, M. (2017). Huddler: Convening stable and familiar crowd teams despite unpredictable availability. In *Proceedings of the 2017 ACM Conference on Computer-Supported Cooperative Work and Social Computing* (pp. 1700–1713). ACM.
- Salk, C. F., Sturm, T., See, L., Fritz, S., & Perger, C. (2016). Assessing quality of volunteer crowdsourcing contributions: Lessons from the Cropland Capture game. *International Journal of Digital Earth*, 9(4), 410–426.
- Satzger, B., Psailer, H., Schall, D., & Dustdar, S. (2013). Auction-based crowdsourcing supporting skill management. *Information Systems*, 38(4), 547–560.
- Schmitz, H., & Lykourantzou, I. (2018). Online sequencing of non-decomposable macrotasks in expert crowdsourcing. *ACM Transactions on Social Computing*, 1(1), 1.

- See, L., Schepaschenko, D., Lesiv, M., McCallum, I., Fritz, S., Comber, A., et al. (2015). Building a hybrid land cover map with crowdsourcing and geographically weighted regression. *ISPRS Journal of Photogrammetry and Remote Sensing*, 103, 48–56.
- Sitkin, S. B., & George, E. (2005). Managerial trust-building through the use of legitimating formal and informal control mechanisms. *International Sociology*, 20(3), 307–338.
- Sorokin, A., Berenson, D., Srinivasa, S. S., & Hebert, M. (2010, October). People helping robots helping people: Crowdsourcing for grasping novel objects. In *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems* (pp. 2117–2122). IEEE.
- Sprugnoli, R., Moretti, G., Fuoli, M., Giuliani, D., Bentivogli, L., Pianta, E., ... & Brugnara, F. (2013, May). Comparing two methods for crowdsourcing speech transcription. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 8116–8120). IEEE.
- Star, S., & Griesemer, J. (1989). Institutional ecology, ‘translations’ and boundary objects: Amateurs and professionals in Berkeley’s Museum of Vertebrate Zoology, 1907-39. *Social Studies of Science*, 19(3), 387–420.
- Stolee, K. T., & Elbaum, S. (2010, September). Exploring the use of crowdsourcing to support empirical studies in software engineering. In *Proceedings of the 2010 ACM-IEEE International Symposium on Empirical Software Engineering and Measurement* (p. 35). ACM.
- Su, H., Deng, J., & Fei-Fei, L. (2012, July). Crowdsourcing annotations for visual object detection. In *Workshops at the Twenty-sixth AAAI Conference on Artificial Intelligence* (Vol. 1, No. 2).
- Tai, L., Chuang, Z., Tao, X., Ming, W., & Jingjing, X. (2011). Quality control of crowdsourcing through workers [sic] experience. In *Proceedings of the ACM SIGIR Workshop on Crowdsourcing for Information Retrieval*.
- Tang, W., & Lease, M. (2011, July). Semi-supervised consensus labeling for crowdsourcing. In *SIGIR 2011 Workshop on Crowdsourcing for Information Retrieval (CIR)* (pp. 1–6).
- Tiwana, A. (2010). Systems development ambidexterity: Explaining the complementary and substitutive roles of formal and informal controls. *Journal of Management Information Systems*, 27(2), 87–126.
- Tran-Thanh, L., Huynh, T. D., Rosenfeld, A., Ramchurn, S. D., & Jennings, N. R. (2014, May). BudgetFix: Budget limited crowdsourcing for interdependent task allocation with quality guarantees. In *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-agent Systems* (pp. 477–484). International Foundation for Autonomous Agents and Multiagent Systems.
- Trompette, P., Chanal, V., & Pelissier, C. (2008, July). Crowdsourcing as a way to access external knowledge for innovation. In *24th EGOS Colloquium*.
- Turner, K. L., & Makhija, M. V. (2006). The role of organizational controls in managing knowledge. *Academy of Management Review*, 31(1), 197–217.
- Ul Hassan, U., Zaveri, A., Marx, E., Curry, E., & Lehmann, J. (2016, November). ACryLIQ: Leveraging DBpedia for adaptive crowdsourcing in linked data quality assessment. In *European Knowledge Acquisition Workshop* (pp. 681–696). Cham: Springer.
- Vempaty, A., Varshney, L. R., & Varshney, P. K. (2014). Reliable crowdsourcing for multi-class labeling using coding theory. *IEEE Journal of Selected Topics in Signal Processing*, 8(4), 667–679.
- Venetis, P., & Garcia-Molina, H. (2012, August). Quality control for comparison microtasks. In *Proceedings of the First International Workshop on Crowdsourcing and Data Mining* (pp. 15–21). ACM.
- Vliegendhart, R., Larson, M., Kofler, C., Eickhoff, C., & Pouwelse, J. (2011, February). Investigating factors influencing crowdsourcing tasks with high imaginative load. In *Proceedings of the Workshop on Crowdsourcing for Search and Data Mining (CSDM) at the Fourth ACM International Conference on Web Search and Data Mining* (pp. 27–30). ACM.
- Wais, P., Lingamneni, S., Cook, D., Fennell, J., Goldenberg, B., Lubarov, D., ... & Simons, H. (2010). Towards building a high-quality workforce with Mechanical Turk. In *Proceedings of Computational Social Science and the Wisdom of Crowds (NIPS)* (pp. 1–5).

- Wang, S., Huang, C. R., Yao, Y., & Chan, A. (2014). Exploring mental lexicon in an efficient and economic way: Crowdsourcing method for linguistic experiments. In *Proceedings of the 4th Workshop on Cognitive Aspects of the Lexicon (CogALex)* (pp. 105–113).
- Weibel, A., Den Hartog, D. N., Gillespie, N., Searle, R., Six, F., & Skinner, D. (2016). How do controls impact employee trust in the employer? *Human Resource Management*, 55(3), 437–462.
- Windeler, J. B., Maruping, L. M., Robert, L. P., & Riemenschneider, C. K. (2015). E-profiles, conflict, and shared understanding in distributed teams. *Journal of the Association for Information Systems*, 16(7), 608.
- Wu, C. C., Chen, K. T., Chang, Y. C., & Lei, C. L. (2013). Crowdsourcing multimedia QoE evaluation: A trusted framework. *IEEE Transactions on Multimedia*, 15(5), 1121–1137.
- Xia, T., Zhang, C., Xie, J., & Li, T. (2012, September). Real-time quality control for crowdsourcing relevance evaluation. In *2012 3rd IEEE International Conference on Network Infrastructure and Digital Content (IC-NIDC)* (pp. 535–539). IEEE.
- Ye, T., You, S., & Robert, L. P. (2017). When does more money work? Examining the role of perceived fairness in pay on the performance quality of crowdworkers. In *Proceedings of the 11th International AAAI Conference on Web and Social Media*.
- You, S., Robert Jr, L. P., & Rieh, S. Y. (2015, April). The appropriation paradox: Benefits and burdens of appropriating collaboration technologies. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems* (pp. 1741–1746). ACM.
- Yung, D., Li, M. L., & Chang, S. (2014). Evolutionary approach for crowdsourcing quality control. *Journal of Visual Languages & Computing*, 25(6), 879–890.
- Zaidan, O. F., & Callison-Burch, C. (2011, June). Crowdsourcing translation: Professional quality from non-professionals. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (Vol. 1, pp. 1220–1229). Association for Computational Linguistics.
- Zhai, H., Lingren, T., Deleger, L., Li, Q., Kaiser, M., Stoutenborough, L., & Solti, I. (2013). Web 2.0-based crowdsourcing for high-quality gold standard development in clinical natural language processing. *Journal of Medical Internet Research*, 15(4).
- Zhang, G., & Chen, H. (2013, October). Quality control for crowdsourcing with spatial and temporal distribution. In *International Conference on Internet and Distributed Computing Systems* (pp. 169–182). Berlin, Heidelberg: Springer.
- Zhang, G., & Chen, H. (2013, December). Quality control of massive data for crowdsourcing in location-based services. In *International Conference on Algorithms and Architectures for Parallel Processing* (pp. 112–121). Cham: Springer.
- Zogaj, S., & Bretschneider, U. (2014). Analyzing governance mechanisms for crowdsourcing information systems: A multiple case analysis. In *Proceedings of the European Conference on Information Systems 2014*.

Chapter 4

Addressing Cooperation Issues in Situated Crowdsourcing



Jorge Goncalves, Simo Hosio, Niels van Berkel and Simon Klakegg

Abstract Situated crowdsourcing has been growing in popularity as an alternative way to collect complex and often creative crowd work. However, previous situated crowdsourcing deployments have not successfully leveraged cooperation possibilities with their audiences, which can improve the data quality of deployed macrotasks. In this chapter, we present three situated crowdsourcing case studies that used different situated technologies and identify the reasons behind their missteps regarding promoting cooperation between workers. Then, based on the identified issues, we propose the design of a novel situated crowdsourcing platform that aims to effectively support cooperation without alienating solo workers. In order to gather insights on our proposed design, we built a prototype platform and evaluated it using a laboratory study with 24 participants. In general, participants were positive about the idea as it provided an easy way to cooperate with friends when completing tasks, while also allowing them to adjust the working environment to their liking. Finally, we conclude by offering insights towards improving cooperation in future situated crowdsourcing deployments and how this can assist in completing macrotasks.

4.1 Introduction

Situated crowdsourcing has emerged as a promising new crowdsourcing paradigm, aimed at providing a complementary means to elicit crowd contributions (Hosio et al. 2014). It entails embedding situated input technologies (e.g. public displays, tablets) in a physical space and leveraging users' serendipitous availability (Müller et al. 2010) or idle time ['cognitive surplus' (Shirky 2010)]. Due to its' characteristics, situated crowdsourcing enables the collection of crowd contributions that can be

J. Goncalves (✉) · N. van Berkel
School of Computing and Information Systems, The University of Melbourne, Parkville, VIC
3010, Australia
e-mail: jorge.goncalves@unimelb.edu.au

S. Hosio · S. Klakegg
Center for Ubiquitous Computing, University of Oulu, Pentti Kaiteran Katu 1, P.O. Box 4500,
90014 Oulu, Finland

challenging to gather with other forms of crowdsourcing (e.g. online). For instance, it allows for targeting of specific individuals' in a certain location (Goncalves et al. 2014a; Heimerl et al. 2012), gathering people's local knowledge on a particular topic (Goncalves et al. 2014b) or reaching an untapped source of potential workers (Hosio et al. 2014). For these reasons, the number of situated crowdsourcing deployments reported in literature is on the rise (e.g. Heimerl et al. 2012; Goncalves et al. 2013, 2016, 2017; Hosio et al. 2014; Huang 2015; Ludwig et al. 2017).

Situated crowdsourcing also opens up opportunities to conduct macrotasks by targeting workers with specific expertise or knowledge. However, macrotasking often involves worker cooperation (Schmitz and Lykourantzou 2018), which is an important challenge with situated crowdsourcing deployments as it can be difficult to promote and/or design for cooperation between the workers. One of the reasons behind this is the inherently public nature of the situated technologies used in these deployments, which has been shown to sometimes lead to disruptive and non-serious behaviours (Kuikkaniemi et al. 2011). Further, while online and mobile crowdsourcing allows each individual to use their own personal device, and facilitate the design of tasks that support cooperation, in situated crowdsourcing there is typically only access to one single device within a specific location. At the same time, situated crowdsourcing deployments in literature have not provided appropriate scaffolding to support cooperation between users, which further exacerbates the aforementioned issues.

In this chapter, we summarise the findings of three situated crowdsourcing deployments using different types of situated technologies (single-purpose large public displays, crowdsourcing kiosks embedded with tablets, and multipurpose public displays) in terms of cooperation between the workers, and discuss the lessons learned. We then propose a novel design for a situated crowdsourcing platform to better support cooperation between workers based on these lessons, which in turn can facilitate the completion of macrotasks. Finally, we present preliminary qualitative results on users' opinions on the prototype design's appropriateness for crowdsourcing and discuss the potential of situated crowdsourcing with regards to the deployment of macrotasks.

4.2 Related Work

4.2.1 Cooperation in Crowdsourcing

Online crowdsourcing platforms have enabled cooperation between workers using computational systems without any limiting spatiotemporal boundaries. However, cooperation is rarely an explicit feature of the work. It is the requesters who must divide, distribute and combine the received work to make it a cohesive whole (Martin et al. 2016). For instance, using *Etherpad* (a lightweight collaborative online notepad), workers from MTurk have been successfully tasked with translating Span-

ish poems into English (Kittur 2010). *Flash Teams* is a framework to coordinate experts from a crowd to perform, e.g. rapid design prototyping or course development (Retelny et al. 2014). They also explore how to create entire organisations consisting of teams with different skill sets which can practically provide output 24 h per day, as the workforce is truly global. Yet another example is *Huddler*, that is used to assemble familiar teams during uncertain availability from MTurk. Huddler (Salehi et al. 2017) provides a thin wrapper where workers wait for other workers to join the ad hoc team before proceeding to complete the actual tasks. Haas et al. presented Argonaut, a framework that improves macrotask-powered work quality using a hierarchical review (Haas et al. 2015).

However, crowdsourcing platforms do not always support cooperation between its users and there is great variation in the extent and nature of the collaboration that occurs (Saxton et al. 2013). Innocentive is a good example of a crowdsourcing platform that only permits partial collaboration in order to safeguard the intellectual property of the task requesters. When its users receive notification about available challenges, they can either tackle it as an individual or with agreed-upon team members available through a confidential Team Project room. Similarly, workers of the platform Upwork can either complete the tasks alone or invite other community members to form a project group.

Beyond cooperation within crowdsourcing platforms, workers of Amazon's Mechanical Turk (i.e. Turkers) have developed elaborate ways to cooperate to identify lucrative tasks or to recreate the social aspects that exist in traditional brick and mortar work (Gray et al. 2016). This is important, as it has been argued that requesters can in some cases benefit from the lack of cooperation amongst workers (Felstiner 2012). Hence, external tools that support cooperation amongst Turkers so they can work together to exert more control over the crowdsourcing market (Martin et al. 2014) are sometimes necessary. Ultimately, Turkers' influence over the platform will depend on the available tools and on how much workers are willing to share their perspectives and actions with others (Martin et al. 2014). Furthermore, a better understanding of how work is actually done can help designers and software engineers who are developing tools to support that work (Gupta et al. 2014). Here, we highlight the challenges of cooperation in situated crowdsourcing and propose a design aimed at providing appropriate scaffolding for cooperation without relying on external tools.

4.2.2 Situated Technologies and Their Use for Crowdsourcing

A key human characteristic that situated technologies, such as displays, can exploit is the fundamental need to explore, to start using pieces of technology rather serendipitously and simply to 'kill time' (Müller et al. 2010). Thus, situating the deployment somewhere people typically have free time is considered beneficial. Furthermore, the

presence of users who are publicly interacting with a deployment draws the attention of passers-by—a phenomenon better known as the *honey-pot effect*—as observed by Brignull and Rogers (2003). The honey-pot effect can be leveraged to increase interactions with a deployment simply by designing for attention and affording the audience to start using the deployment (Hosio et al. 2016). However, deployments often tend to support only one simultaneous user, and the honey-pot effect leads to queuing, which can be detrimental to the overall experience.

Furthermore, situated technology deployments are often used by groups of users (Hosio et al. 2016). However, while using technologies in groups of people is fun and entices interaction, social awkwardness has also been documented in such situations. For example, the space around a public deployment can be perceived as a proverbial stage where the audience is watching the user (Kuikkaniemi et al. 2011). The group members also sometimes conflict with each other when using a shared deployment. For example, Peltonen et al. studied social group interactions in an urban city area using a public display as an intervention (Peltonen et al. 2008). They document in detail how the presence of users invited new interactions to take place with the deployment, and how the presence of others often leads to conflicts and tensions in the personal spaces of users.

Situated technologies have certain desired characteristics for crowdsourcing, such as low barrier of entry for people who would not otherwise engage in crowdsourcing or targeting a specific group of wanted participants (Goncalves et al. 2013; Hosio et al. 2014). However, the aforementioned issues with situated technologies also hinder their potential for crowdsourcing purposes, and with our work, we seek to pinpoint and offer solutions to identified cooperation challenges.

4.2.2.1 Situated Crowdsourcing Deployments

Crowdsourcing using situated technologies is becoming more and more feasible as the number of installations grows. A recent example of a situated crowdsourcing deployment is *Umati*, an augmented vending machine used to explore CommunitySourcing (Heimerl et al. 2012). *Umati* dispatched edible goods such as snacks and chocolate in exchange for labour that could only be completed accurately by local workers. *Bazaar*, by Hosio et al., investigated how an economic market model applies in situated settings, concluding that the supply of labour can indeed be controlled with alternating the rewards also in situated task markets (Hosio et al. 2014). The same platform was later used to explore the collection of subjective and local data as well (Goncalves et al. 2017). Two more recent examples include *CrowdFeedBack* and *CrowdButton* that together focus on sustaining the uptake and quality of unpaid crowdsourcing contributions (Huang 2015). As a final example, *City-Share* facilitates efficient communication between official emergency personnel and volunteers in disaster zones by using public displays as communication hubs (Ludwig et al. 2017). With the rise of situated crowdsourcing deployments, Huang et al. (2017) proposed a genetic model inspired by the MIT's model on collective intelligence (Malone et al.

2010), aimed at helping researchers in this area by identifying important contextual aspects for user contributions in situated crowdsourcing systems.

Despite the much-explored potential, situated crowdsourcing deployments are inherently limited by both scale and reach. Contrary to traditional online crowdsourcing, where a deployment can potentially reach millions of users (Ipeirotis and Gabrilovich 2014) who contribute using their own familiar devices anywhere, in situated settings the workers typically complete tasks using devices deployed by third parties as parts of the fixed environment. For this reason, researchers consider situated crowdsourcing more as an alternative, or different means of eliciting crowd contributions, rather than a replacement or competitor of online crowdsourcing (Goncalves et al. 2013).

4.3 Case Studies

Next, we summarise the findings regarding cooperation between workers of three separate situated crowdsourcing deployments that used different types of situated technologies, namely: (1) single-purpose public displays, (2) multipurpose public displays and (3) kiosks embedded with tablets.

4.3.1 Case Study 1 (C1): Crowdsourcing Malaria Detection

Our first case study entailed using four 46" single-purpose public displays (Fig. 4.1) to crowdsource malaria detection. The task entailed asking workers to count malaria-infected blood cells on images of a petri dish generated algorithmically while comparing different motivational approaches. More details on this deployment can be seen in Goncalves et al. (2013).

What makes this particular deployment unique is that we recorded all interactions with one public display, and thus we were able to observe participants' attitudes and social context when completing tasks. Our video recordings consisted of 123 distinct instances of interaction and based on content analysis using open and axial coding we identified different emerging themes of behaviour. As reported in Goncalves et al. (2013), this analysis confirmed instances of the behaviours that we initially noted in our in situ observations, but also revealed several new behaviours that people exhibited when using the display. The six identified behaviours were:

- **Ignorer:** passers-by that ignored the display, exhibiting what is often referred to as display blindness (Müller et al. 2009), and
- **Unlocker:** those that actually unlocked the screen but completed no tasks. These account for the high number of curiosity clicks mentioned previously.
- **Herder:** individuals would approach the display with a group of people, complete some tasks and then leave with the group. The other members would adopt a



Fig. 4.1 Example of one of the single-purpose public displays used in this deployment

passive position behind the herder, in a way that suggested they were not applying social pressure but rather observing,

- **Loner:** individuals that approached the display alone and typically spent more time than others completing tasks.
- **Attractor:** attracted others to join them on the display, commonly referred to as the honey-pot effect (Brignull and Rogers 2003), and complete tasks jointly.
- **Repeller:** applied social pressure to try to make the worker leave the display. Instances of repellers also happened when groups of two or more people approached the display.

A visual representation of each of these behaviours can be seen in Fig. 4.2. Overall, analysis of the work conducted by each group of workers showed that sole users, dubbed as *loners*, spent more time completing tasks. More specifically, loners completed on average a higher number of tasks ($M = 4.91$) when compared to the other groups: attractors ($M = 3.71$), herders ($M = 3.43$) and finally repellers ($M = 1.29$). A Kruskal–Wallis test showed that there was a significant difference in average number of tasks completed between the different behaviours ($\chi^2(4) = 22.18, p < 0.01$). Post hoc analysis using the Mann–Whitney tests showed that there was only a significant difference between loners and repellers in terms of average number of completed tasks ($U = 26.04, p < 0.01$). As for accuracy, a Kruskal–Wallis test showed that there was no significant difference in accuracy between the different behaviours ($\chi^2(4) =$

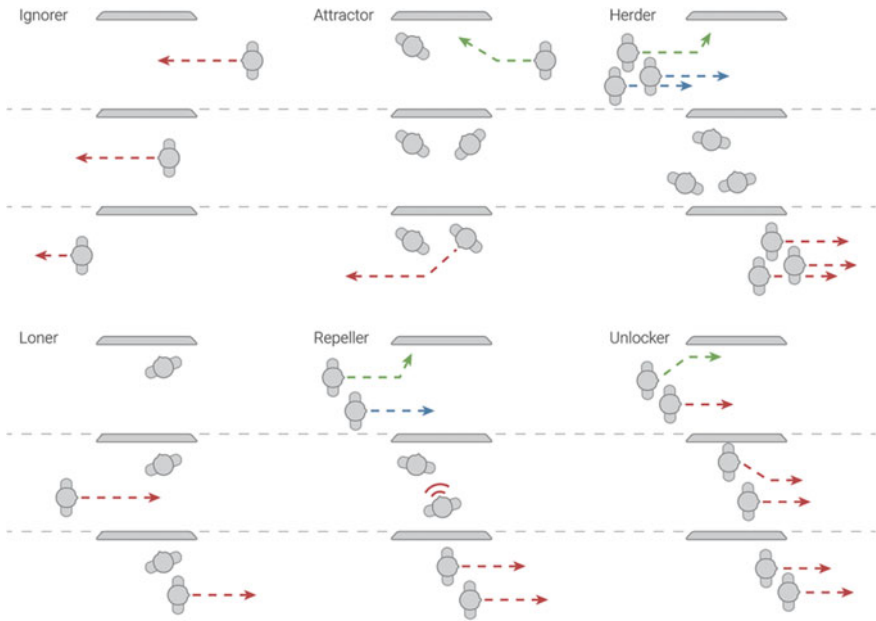


Fig. 4.2 The six identified behaviours in this deployment

7.99, $p = 0.09$). These results suggest that while not having a significant impact on accuracy, groups of workers spent significantly less time completing the tasks.

These results can be seen as problematic as situated technologies naturally invite groups of people to engage with them or have people join those already engaging with the technology during interaction [known as the honey-pot effect (Brignull and Rogers 2003)]. Furthermore, previous work has suggested that this behaviour is effective in combating feelings of self-consciousness felt by a solo user when engaging with public technologies (Kuikkaniemi et al. 2011). Unfortunately, the latter two behaviours (attractor and repeller) ultimately led to a disturbance and delay in the completion of the tasks. Here, the workers were not encouraged to perform well, but instead engage in performative acts (Hosio et al. 2015) resulting in non-serious completion of tasks. In fact, previous work suggests that in some cases the engagement with these interactive public artefacts emerges only when the overall social context provides a ‘license to play’ (Jurmu et al. 2014). In the case of playful applications or games, this does not matter and can even act as a catalyst to use (Kuikkaniemi et al. 2011), but for crowdsourcing purposes where meaningful data is being collected from the public, it is important to provide appropriate scaffolding for group use. If additional individuals feel like they are not able to contribute meaningfully, then this will ultimately lead to them disturbing those that are engaging with the platform. While this particular case study involved microtasks, we argue that our findings generalise to the completion of macrotasks in a situated crowdsourcing deployment (i.e. similar context).

4.3.2 Case Study 2 (C2): Crowdsourcing Public Opinion

Our second case study deals with a public, large-scale in-the-wild deployment at the heart of downtown Oulu, in Finland. Collecting and analysing city-scale feedback from individual citizens is one way of crowdsourcing the public opinion (Hosio et al. 2015). In this case study, we used a grid of interactive large public displays, *UBI-hotspots* (Hosio et al. 2016), to elicit civic feedback from the young (Hosio et al. 2015). More specifically, we deployed a photo booth application that was paired with social media to enable a two-way communication channel between citizens and officials.

UBI-hotspots (as seen in Fig. 4.3) are large displays deployed in pivotal locations in Oulu. These displays host several applications, i.e. they are *multipurpose* (Hosio et al. 2013). This ‘battle’, where every application has several contenders for user attention, led to designing the application as playful in the first place. At the time, we reasoned it is fair to anticipate the younger generations to be drawn into gamified concepts rather than ‘boring’ civic affairs. The key design choices in addition to playfulness were to exploit the attractiveness of public technologies in general (Müller et al. 2010) and to extend interaction capabilities by leveraging social media.

In terms of the original goal, i.e. providing a useful two-way discussion channel between the young and the city youth affairs department, the six months-long deployment turned out to be a quite the fiasco. While the volume of submissions, or feedback items, was fairly satisfactory (425 unique submissions), it soon became painfully clear that practically none of them had anything to do with the original goal of the deployment. No feedback was being crowdsourced, and the deployed system was used for toying around and for taking snapshots for the sake of having fun. A representative sample of the submitted entries can be seen in Fig. 4.4.

In hindsight and regards to situated crowdsourcing, we identify an important aspect worth considering in the design stage. While playful design elements that are often praised in related literature as good ways to elicit engagement, it backfired in



Fig. 4.3 Example of one of the multipurpose public displays used in this deployment



Fig. 4.4 Citizen feedback submitted through our crowdsourcing platform. Top row: teenagers playing with an energy drink can, tourists taking pictures with the deployment, teens acting to the camera. Bottom row: groups of people posing for the camera

this type of ‘serious’ application. Granted, the young did enjoy using the application, and at times spent several minutes with it in order to create beautiful sequences of pictures, but for the ‘wrong’ purpose. One can learn a lot from human behaviour such as demonstrated in the submissions (Hosio et al. 2015), but this takes a lot of effort and does not necessarily answer to the original needs of the deployment. Providing feedback to the city was simply a lesser motive than having fun with the tech just to take ‘funny’ pictures. That being said, the big screens we used were clearly suitable for ad hoc cooperation to take place: the large screens were used as toys to play with, and especially the camera was seen as a motivator to engage with the application. In that sense, designing for playfulness that channels the energy and exploration to the intended direction can be beneficial.

4.3.3 Case Study 3 (C3): Situated Crowdsourcing Market

Our third and final case study entailed the development and deployment of a situated crowdsourcing market, called *Bazaar*, using multiple public kiosks embedded with tablets deployed in different locations (Fig. 4.5). The platform enabled users to create accounts, earn virtual currency by completing a number of different types of tasks (e.g. sentiment analysis, image labelling) and exchange earned currency for rewards (e.g. money, movie tickets, coffee vouchers, etc.). More details on this deployment can be seen in Hosio et al. (2014).

Here, one of our intentions was to provide a more private means to complete crowd work and mitigate any self-conscious issues when engaging with public technology. However, as a direct result of the smaller screen estate, collaborative work between

Fig. 4.5 Example of one of the kiosks used during this deployment



the workers became more challenging. In fact, during our interviews, several users of *Bazaar* reported wanting to work towards a common prize or simply help a friend complete the given tasks. Unfortunately, the platform did not support this, which in some cases resulted in the workers quitting the platform altogether. Those with friends that continued using the platform found alternative ways to achieve their goals, such as sharing accounts or, more commonly, working separately in different locations instead of cooperating in a meaningful way. Given the distance between the different kiosks, this solution proved to be rather non-ideal removing any social aspects from conducting the crowdsourcing work. Several groups of workers completed tasks until they all achieved a certain goal (i.e. each person getting enough virtual currency to get a movie ticket), and then stopped using the platform. In case a similar deployment was to be conducted in the future to support cooperation between workers, then a collocated solution could prove more efficient in attracting and engaging workers with the platform. Finally, the design was deemed as not enough customizable in terms of ergonomic factors: workers wanted to adjust the height or even the angle of the display, as in many cases the sun or other lights were reflecting from the embedded tablet's surface.

4.3.4 *Summary of Identified Issues*

First, one major pitfall in our presented case studies (and other situated crowdsourcing deployments), is that they did not allow more than one person to directly engage with the tasks simultaneously. This can ultimately lead to a disturbance that will affect the worker engaging with the tasks (as seen in C1 and C2). One potential solution is to allow additional people to use their own personal devices to contribute to the crowdsourcing task (e.g. mobile phone). However, it is challenging to provide reliable runtime assembly of multi-device ecologies (Heikkinen et al. 2014;

Weißker et al. 2016), and without seamless interactivity, workers can quickly lose interest. Furthermore, previous work has shown that adding additional barriers to participation can significantly hinder the likelihood people will engage with a situated crowdsourcing platform (Goncalves et al. 2013). Hence, offering a simple and rapid solution to enable cooperation in these settings is crucial.

Second, the type of situated technology will significantly affect what kind of work can be conducted and how cooperation should be supported. Given the added control, better usability and more private crowdsourcing experience (i.e. smaller screens meant that others could not see what a worker was doing) of situated kiosks, we argue that they are better suited to support cooperation in situated crowdsourcing deployments. However, while the experiment reported in C3 did indeed offer these benefits, it also restricted even further any possibility for cooperation. Several workers that interacted with the platform were eager to cooperate with others, and ended up taking alternative routes to achieve this goal. Thus, we argue that a multiple input and collocated solution would trump multi-location deployments (such as the one reported in C3) when cooperation between workers is desired.

Finally, while designing with playfulness in mind has been showed in the past to be highly successful in engaging users with situated technologies (Kuikkaniemi et al. 2011), in C2 it was highly detrimental to the original intent of the experiment: to crowdsource public opinion on a specific matter. This is not to say that performing tasks in a situated crowdsourcing environment should not be enjoyable, but that the design should minimise as much as possible appropriation of the technology by workers for different purposes than originally intended.

4.4 Proposed Design

In order to mitigate the issues identified in our case studies, we designed and constructed a situated crowdsourcing table with three attached tablets (Fig. 4.6). The design of this table was informed by the findings reported in our case studies, as well as years of experience conducting situated crowdsourcing experiments. We settled for three tablets as we rarely saw larger groups engage with the display in C1 and also because it allows for a few to few ecosystems that enables natural interaction between the workers to occur (Terrenghi et al. 2009). This is not to say that the platform would enforce three simultaneous workers, but allow for up to three workers to interact with the available crowdsourcing tasks. The proposed design also enables solo workers to complete tasks if they so choose, including simultaneous solo workers that do not wish to cooperate. It would then be up to the task requester and/or designer to decide which tasks available on the platform would support cooperation and which would not. When designing for cooperation this could be achieved directly on the interface (e.g. workers interact with the same task simultaneously to solve it), or indirectly by simply encouraging communication between the workers (e.g. each worker interacts with different subtasks of a larger task). For instances of direct cooperation, assigning a leader may be necessary to ensure high-quality task

Fig. 4.6 Situated crowdsourcing table with three attached tablets



completion, and to coordinate and submit the work, as suggested in Retelny et al. (2014). We also anticipate having a responsive leader in each session to, at least in some cases, reduce the amount of non-serious behaviour in other workers.

Furthermore, the tablets are placed within a special enclosure to prevent appropriation of the technology as seen in C2 (e.g. power button is inaccessible). Furthermore, a registration process required before completing any tasks can filter out non-serious individuals (Hosio et al. 2014). The enclosure rests upon a hinge, allowing workers to reposition the tablet vertically [adapting the visual angle as suggested in Terrenghi et al. (2009)] and potentially show to the other workers what is currently on their screen. In addition, the enclosure allows the worker to rotate the tablet as deemed necessary. We opted for a round table to promote conversation and cooperation between workers currently working on the same task, as seen in Shen et al. (2003). The table's height is also adjustable to cater to a more diverse set of potential workers and promote inclusivity. While there will be issues when workers of very different heights engage with the platform, we argue that this design is still more inclusive than past situated crowdsourcing deployments reported in the literature that uses a technology with a predefined and non-changeable height. As an example, the kiosks presented in C3 did not allow workers to adjust the height of the screen or the visual angle, making for a non-ideal working experience for some workers. A summary of the identified issues and the design choices aimed at solving them can be seen in Table 4.1.

4.4.1 Interview Procedure and Method

We recruited 24 participants from mailing lists in our university and social media. Recruited participants were from several different study areas such as computer science, biomedical engineering, biology, education, and product management. In our usability lab, we showed the participants the table and allowed them to directly

Table 4.1 Identified issues from our case studies and solutions offered by our proposed design

Identified issue	Design choices
Idle friends disrupting others' work (C1, C2)	Multiple collocated devices
Excessive appropriation (C2)	Enclosure that hides certain functions. Appointing a responsible leader, requiring registration
Cooperation not supported (C1, C3)	Generic platform that allows, but does not enforce cooperation in tasks
Physical limitations, work ergonomomy (C3)	Adjustable table design (height, device angle, rotates)

interact with it. We conducted the semi-structured interviews in groups of 3 with each one lasting around 15 min. During the interviews, we asked their opinions regarding the design of the table, what tasks would work well or not with the setup, and the benefits and drawbacks the proposed design would have over other situated technologies (e.g. large public display) for completing macrotasks. Participants were given a movie voucher for their participation.

We used thematic analysis to explore our qualitative data. Thematic analysis is 'a method for identifying, analysing, and reporting patterns (themes) within data' and is commonly applied in qualitative research (Braun and Clarke 2006). First, we extracted the qualitative data from our responses, and focused on discovering different themes. We then wrote simple descriptive notes on these themes and discussed them. Since our research is largely exploratory without a theoretical framework about designing for cooperation in situated crowdsourcing, our coding process was inductive. Codes emerged and were selected through an iterative process and discussion between the coders.

4.4.2 Results

4.4.2.1 Input Mechanisms

Participants expressed that the number of available tablets would most likely be sufficient in most cases, but at the same time could see how a higher number of tablets could sometimes be useful. By offering several simultaneous input mechanisms it is more likely that present individuals express their opinions when compared to the typical single input mechanism platform reported in the majority of situated crowdsourcing deployments. This was seen as particularly useful in the case of macrotasks where simultaneous input could facilitate the completion of the tasks. This effectively breaks these tasks into microtasks, which has been shown to result in higher quality outcomes and a better experience that can reduce the impact of interruptions (Cheng et al. 2015).

If this happens on one screen—maybe some people might not express their opinion. So if we have three tablets we can be sure that everyone mentions their opinion. (P12)

Definitely with more complex tasks having separate inputs is great, instead of everyone trying to chime in on the same screen. Less confusion and more likely that everyone contributes. (P17)

4.4.2.2 Table Design

Furthermore, one group of participants appreciated the privacy aspects of our design, stating that it would be much more awkward to complete tasks on a larger display. This is in line with previous work on public displays that report feelings of self-consciousness when interacting with a large display in public areas (Kuikkaniemi et al. 2011).

I would feel awkward or embarrassed when doing it on a larger screen, so I prefer smaller screens for this. (P02)

This is of particular importance when completing more sensitive tasks that workers might, in general, be less comfortable completing, and may even prefer completing them alone, a possibility that is also possible in our proposed design.

In addition, several participants identified the repositioning features of the tablet enclosures as a beneficial way to quickly show others what is on their screen, thus supporting cooperation between the workers. Finally, participants appreciated the ability to rotate the tablet to a more comfortable position as typically situated crowdsourcing deployments can be quite tiring when completing tasks for an extended period of time.

4.4.2.3 Collocated Interaction

While all situated crowdsourcing deployments have elements of collocated interaction, participants reported that our design could further facilitate these interactions. The round design of the table and closeness of each tablet was seen as an important enabler for better communication between the workers. Unlike situated crowdsourcing deployments that use public displays and have workers stand side by side, our design positions workers to be face-to-face facilitating interaction.

Communication is better, you can see the faces, impressions, and everything. (P05)

I like the fact that it is a round table, because you can see each other faces. It facilitates conversation, a big screen would be worse. You cannot experience the feelings of people etcetera. I also kind of like that everyone has their own screen. (P08)

In addition, workers can more easily identify if others are unsure of their answers or not contributing sufficiently to the tasks.

It can help as you can see the body language or if someone is a little bit shy or not saying things. (P22)

4.4.2.4 Task Suitability

During the interviews, some participants expressed the suitability of different types of tasks to the proposed design. For instance, visual search tasks (e.g. finding a certain object in an image) would benefit from all workers interacting on the same screen.

If we could have one big screen for searching, that would be good. Just one screen for all of us. (P13)

If we all have one big screen, it is easier to see what everyone is looking at - or are pointing. (P20)

This can be explained by the fact that such tasks are objective and have only 1 correct answer. By having all participants look at a single screen will lead to faster completion times, and therefore a more efficient workflow. However, for cooperation in most types of tasks, participants agreed that the proposed design would be more advantageous over a larger public display. For instance, in more subjective tasks workers are able to discuss and potentially annotate parts of a task without disturbing the view of the others.

4.4.3 Lessons Learned

In this section, we summarise the lessons learned through the design and evaluation of our situated crowdsourcing platform. Situated crowdsourcing enables crowd work that requires local knowledge or that benefits from face-to-face interactions, tasks that are challenging to complete with online crowdsourcing, so appropriately supporting this collaboration is crucial. Participants of our user study praised the approach for allowing easy collocated cooperation between workers and adjusting the work environment to their specifications. In addition, the use of tablets over large public displays was mostly seen as beneficial in preserving privacy as well as promoting discussion between the different workers. Moreover, the better usability of these devices can facilitate the completion of macrotasks, which can be challenging to complete in a situated crowdsourcing setting due to task complexity and the increased likelihood that workers can be distracted by the surrounding environment. Furthermore, while the proposed design may not be ideal for cooperation in every type of tasks, it was considered as being an effective approach to provide, in most cases, sufficient scaffolding for cooperation between situated crowdsourcing workers. Finally, while completing macrotasks using our design is likely to result in longer completion times, it is also likely to result in higher quality outcomes and a better experience as it breaks these tasks into more manageable microtasks (Cheng et al. 2015).

In general, it is crucial for researchers to conceptualise new forms of crowd work that go beyond simple and independent tasks that are common today in many crowdsourcing platforms (Kittur et al. 2013). In the case of situated crowdsourcing, allowing and supporting cooperation between collocated workers presents itself as an

important research direction, as macrotasking often involves worker cooperation (Schmitz and Lykourantzou 2018). In that sense, our design was considered by our participants as a positive step towards effective cooperation in situated crowdsourcing settings, as it has the necessary characteristics to facilitate conducting work in a more challenging setting when compared to online crowdsourcing.

4.5 Conclusion and Future Work

Previous deployments in situated crowdsourcing leveraged little or no cooperation between the anticipated workers, thus making it challenging to deploy complex tasks. We argue that this is caused not by an inherent limitation of situated crowdsourcing, but instead it is due to the fact that these deployments did not provide appropriate scaffolding to support said cooperation. With this chapter, we identify specific challenges and flaws in design that have led to this potential shortcoming, in order to inform researchers interested in conducting situated crowdsourcing experiments. Namely, lack of support for several simultaneous workers, inefficient distribution of input mechanisms, design that allowed for appropriation, among others were identified as important challenges that should be considered when designing situated crowdsourcing experiments that support cooperation between workers. Taking these identified challenges into consideration, we then proposed our own design of a situated crowdsourcing platform that facilitates cooperation between workers, and therefore, the completion of relevant macrotasks.

In the future, we hope to implement and evaluate a situated crowdsourcing market that leverages the table design proposed in this chapter. This would entail designing different crowdsourcing tasks for both solo and groups of workers, and conducting an in-the-wild deployment. Ultimately, we argue that it is important to develop new situated crowdsourcing ecologies that support, not enforce, cooperation between workers engaging with the platform and we believe our work presents an important first step towards this goal.

Acknowledgements This work is partially funded by the Academy of Finland (286386-CPDSS, 285459-iSCIENCE), the European Commission (Grant 6AIKA-A71143-AKAI), and Marie Skłodowska-Curie Actions (645706-GRAGE).

References

- Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77–101.
- Brignull, H., & Rogers, Y. (2003). Enticing people to interact with large public displays in public spaces. In M. Rauterberg, M. Menozzi, & J. Wesson (Eds.), *Proceedings of 9th IFIP TC13 International Conference on Human-Computer Interaction, INTERACT '03, Zurich, Switzerland, 1–5 September 2003* (pp. 17–24). Amsterdam: IOS Press.

- Cheng, J., Teevan, J., Iqbal, S. T., & Bernstein, M. S. (2015). Break it down: A comparison of macro- and microtasks. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)* (pp. 4061–4064). ACM Press.
- Felstiner, A. (2012). The weakness of crowds. *Crowds and Clouds*. Retrieved March 15, 2017, from <http://limn.it/the-weakness-of-crowds/>.
- Goncalves, J., Ferreira, D., Hosio, S., Liu, Y., Rogstadius, J., Kukka, H., et al. (2013). Crowdsourcing on the spot: Altruistic use of public displays, feasibility, performance, and behaviours. In J. F. Canny, M. Langheinrich, & J. Rekimoto (Eds.), *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing, UbiComp '13*, Zurich, Switzerland, 8–12 September 2013 (pp. 753–762). New York: ACM Press.
- Goncalves, J., Pandab, P., Ferreira, D., Ghahramani, M., Zhao, G., & Kostakos, V. (2014a). Projective testing of diurnal collective emotion. In J. Kientz, J. Scott, & J. Song (Eds.), *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing, UbiComp '14*, Seattle, USA, 13–17 September 2014 (pp. 487–497). New York: ACM Press.
- Goncalves, J., Hosio, S., Ferreira, D., & Kostakos, V. (2014b). Game of words: Tagging places through crowdsourcing on public displays. In C. Neustaedter, S. Bardzell, & E. Paulos (Eds.), *Proceedings of the 2014 Conference on Designing Interactive Systems, DIS '14*, Vancouver, Canada, 21–25 June 2014 (pp. 705–714). New York: ACM Press.
- Goncalves, J., Kukka, H., Sanchez, I., & Kostakos, V. (2016). Crowdsourcing queue estimations in situ. In P. Bjørn & J. Konstan (Eds.), *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work and Social Computing, CSCW '16*, San Francisco, USA, 27 February–2 March 2016 (pp. 1040–1051). New York: ACM Press.
- Goncalves, J., Hosio, S., & Kostakos, V. (2017). Eliciting structured knowledge from situated crowd markets. *ACM Transactions on Internet Technology*, 17(2), Article 14.
- Gray, M. L., Suri, S., Ali, S. S., & Kulkarni, D. (2016). The crowd is a collaborative network. In P. Bjørn & J. Konstan (Eds.), *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work and Social Computing, CSCW '16*, San Francisco, USA, 27 February–2 March 2016 (pp. 134–147). New York: ACM Press.
- Gupta, N., Martin, D., Hanrahan, B. V., & O'Neill, J. (2014). Turk-life in India. In D. W. McDonald & P. Bjørn (Eds.), *Proceedings of the 18th International Conference on Supporting Group Work, GROUP '14*, Sanibel Island, USA, 9–12 November 2014 (pp. 1–11). New York: ACM Press.
- Haas, D., Ansel, J., Gu, L., & Marcus, A. (2015, August). Argonaut: Macrotask crowdsourcing for complex data processing. In *Proceedings of the VLDB Endow* (Vol. 8, no. 12, pp. 1642–1653). <http://dx.doi.org/10.14778/2824032.2824062>.
- Heikkinen, T., Goncalves, J., Kostakos, V., Elhart, I., & Ojala, T. (2014). Tandem browsing toolkit: Distributed multi-display interfaces with web technologies. In A. Quigley (Ed.), *Proceedings of the International Symposium on Pervasive Displays, PerDis '14*, Copenhagen, Denmark, 3–4 June 2014 (pp. 142–147). New York: ACM Press.
- Heimerl, K., Gawalt, B., Chen, K., Parikh, T., & Hartmann, B. (2012). Community sourcing: Engaging local crowds to perform expert work via physical kiosks. In H. Chi & K. Höök (Eds.), *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '12*, Austin, USA, 5–10 May 2012 (pp. 1539–1548). New York: ACM Press.
- Hosio, S., Goncalves, J., & Kostakos, V. (2013). Application discoverability on multipurpose public displays: Popularity comes at a price. In *Proceedings of International Symposium on Pervasive Displays 2013 (PerDis '13)* (pp. 31–36).
- Hosio, S., Goncalves, J., Kostakos, V., & Riecki, J. (2015a). Crowdsourcing public opinion using urban pervasive technologies: Lessons from real-life experiments in Oulu. *Policy & Internet*, 7(2), 203–222.
- Hosio, S., Goncalves, J., Lehdonvirta, V., Ferreira, D., & Kostakos, V. (2014). Situated crowdsourcing using a market model. In M. Dontcheva & D. Wigdor (Eds.), *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology, UIST '14*, Honolulu, USA, 5–8 October 2014 (pp. 55–64). New York: ACM Press.

- Hosio, S., Harper, R., O'Hara, K., Goncalves, J., & Kostakos, V. (2015). Life through the lens: A qualitative investigation of human behaviour with an urban photography service. In P. Olivier & D. Foster (Eds.), *Proceedings of the 2015 British HCI Conference, British HCI '15*, Lincoln, United Kingdom, 13–17 July 2015 (pp. 157–164). New York: ACM Press.
- Hosio, S., Kukka, H., Goncalves, J., Kostakos, V., & Ojala, T. (2016). Toward meaningful engagement with pervasive displays. *IEEE Pervasive Computing*, 15(3), 24–31.
- Huan, Y., Shema, A., & Xia, H. (2017). A proposed genome of mobile and situated crowdsourcing and its design implications for encouraging contributions. *International Journal of Human-Computer Studies*, 102, 69–80.
- Huang, Y.-C. (2015). Designing a micro-volunteering platform for situated crowdsourcing. In L. Cioffi & D. McDonald (Eds.), *Proceedings of the 19th ACM Conference Companion on Computer-Supported Cooperative Work and Social Computing, CSCW '15*, Vancouver, Canada, 14–18 March 2015 (pp. 73–76). New York: ACM Press.
- Ipeirotis, P. G., & Gabrielovich, E. (2014). Quiz: Targeted crowdsourcing with a billion (potential) users. In A. Broder, K. Shim, & T. Suel (Eds.), *Proceedings of the 23rd International Conference on World Wide Web, WWW '14*, Seoul, South Korea, 7–11 April 2014 (pp. 143–154). New York: ACM Press.
- Jurmu, M., Goncalves, J., Riekkilä, J., & Ojala, T. (2014). Exploring use and appropriation of a non-moderated community display. In S. W. Loke, L. Kulik, & E. Pitoura (Eds.), *Proceedings of the 13th International Conference on Mobile and Ubiquitous Multimedia, MUM '14*, Melbourne, Australia, 25–28 November 2014 (pp. 107–115). New York: ACM Press.
- Kittur, A. (2010). Crowdsourcing, collaboration and creativity. *XRDS: Crossroads, The ACM Magazine for Students*, 17(2), 22–26.
- Kittur, A., Nickerson, J. V., Bernstein, M., Gerber, E., Shaw, A., Zimmerman, J., et al. (2013). The future of crowd work. In C. Lampe & L. Terveen (Eds.), *Proceedings of the 2013 Conference on Computer-Supported Cooperative Work and Social Computing, CSCW '13*, San Antonio, USA, 23–27 February 2013 (pp. 1301–1318). New York: ACM Press.
- Kuikkaniemi, K., Jacucci, G., Turpeinen, M., Hoggan, E., & Müller, J. (2011). From space to stage: How interactive screens will change urban life. *Computer*, 44(6), 40–47.
- Ludwig, T., Kotthaus, C., Reuter, C., van Dongen, S., & Pipek, V. (2017). Situated crowdsourcing during disasters: Managing the tasks of spontaneous volunteers through public displays. *International Journal of Human-Computer Studies*, 102, 103–121.
- Malone, T. W., Laubacher, R., & Dellarocas, C. (2010). The collective intelligence genome. *MIT Sloan Management Review*, 51(3), 21–31.
- Martin, D., Hanrahan, B. V., O'Neill, J., & Gupta, N. (2014). Being a turker. In M. R. Morris, & M. Reddy (Eds.), *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work and Social Computing, CSCW '14*, Baltimore, USA, 15–19 February 2013 (pp. 224–235). New York: ACM Press.
- Martin, D., O'Neill, J., Gupta, N., & Hanrahan, B. V. (2016). Turking in a global labour market. *Computer Supported Cooperative Work*, 25(1), 39–77.
- Müller, J., Alt, F., Michelis, D., & Schmidt, A. (2010). Requirements and design space for interactive public displays. In Smeulders, A. (Ed.), *Proceedings of the 18th ACM International Conference on Multimedia, MM '10*, Firenze, Italy, 25–29 October 2010 (pp. 1285–1294). New York: ACM Press.
- Müller, J., Wilmsmann, D., Exeler, J., Buzeck, M., Schmidt, A., Jay, T., et al. (2009). Display blindness: The effect of expectations on attention towards digital signage. In H. Tokuda, M. Beigl, A. Friday, A. J. Brush, & Y. Tobe (Eds.), *Proceedings of the International Conference on Pervasive Computing, Pervasive '09*, Nara, Japan, 17–20 May 2010 (pp. 1–8). Berlin, Heidelberg: Springer.
- Peltonen, P., Kurvinen, E., Salovaara, A., Jacucci, G., Ilmonen, T., Evans, J., et al. (2008). It's mine, don't touch!: Interactions at a large multi-touch display. In D. Tan (Ed.), *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '08*, Florence, Italy, 5–10 April 2008 (pp. 1285–1294). New York: ACM Press.

- Retelny, D., Robaszekiewicz, S., To, A., Lasecki, W. S., Patel, J., Rahmati, N., et al. (2014). Expert crowdsourcing with flash teams. In M. Dontcheva & D. Wigdor (Eds.), *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology, UIST '14*, Honolulu, USA, 5–8 October 2014 (pp. 75–85). New York: ACM Press.
- Salehi, N., McCabe, A., Valentine, M., & Bernstein, M. (2017). Huddler: Convening stable and familiar crowd teams despite unpredictable availability. In L. Barkhuus, M. Borges, & W. Kellogg (Eds.), *Proceedings of the 2017 ACM Conference on Computer-Supported Cooperative Work and Social Computing, CSCW '17*, San Francisco, USA, 25 February–1 March 2017 (pp. 1700–1713). New York: ACM Press.
- Saxton, G. D., Onook, O., & Kishore, R. (2013). Rules of crowdsourcing: Models, issues, and systems of control. *Information Systems Management*, 30(1), 2–20.
- Schmitz, H., & Lykourantzou, I. (2018, January). Online sequencing of non-decomposable macro-tasks in expert crowdsourcing. *ACM Transactions on Social Computing*, 1, 1, Article 1, 33 pp. <https://doi.org/10.1145/3140459>.
- Shen, C., Everitt, K., & Ryall, K. (2003). UbiTable: Impromptu face-to-face collaboration on horizontal interactive surfaces. In A. K. Dey, A. Schimdt, & J. F. McCarthy (Eds.), *Proceedings of the International Conference on Ubiquitous Computing, UbiComp '03*, Seattle, USA, 12–15 October 2003 (pp. 281–288). Berlin, Heidelberg: Springer.
- Shirky, C. (2010). *Cognitive surplus: How technology makes consumers into collaborators*. Penguin.
- Terrenghi, L., Quigley, A., & Dix, A. (2009). A taxonomy for and analysis of multi-person-display ecosystems. *Personal and Ubiquitous Computing*, 13(8), 583–598.
- Weißker, T., Berst, A., Hartmann, J., & Echtler, F. (2016). The massive mobile multiuser framework: Enabling ad-hoc realtime interaction on public displays with mobile devices. In J. Muller & N. Memarovic (Eds.), *Proceedings of the 5th ACM International Symposium on Pervasive Displays, PerDis '16*, Oulu, Finland, 20–22 June 2016 (pp. 168–174). New York: ACM Press.

Part II
The Role of AI and Experts

Chapter 5

Hybrid Machine-Crowd Interaction for Handling Complexity: Steps Toward a Scaffolding Design Framework



António Correia, Shoaib Jameel, Hugo Paredes, Benjamim Fonseca
and Daniel Schneider

Abstract Much research attention on crowd work is paid to the development of solutions for enhancing microtask crowdsourcing settings. Although decomposing difficult problems into microtasks is appropriate for many situations, several problems are non-decomposable and require high levels of coordination among crowd workers. In this chapter, we aim to gain a better understanding of the macrotask crowdsourcing problem and the integration of crowd-AI mechanisms for solving complex tasks distributed across expert crowds and machines. We also explore some design implications of macrotask crowdsourcing systems taking into account their scaling abilities to support complex work in science.

5.1 Introduction

In recent years, we have seen a flourishing of crowd-powered systems intended to support computer-hard tasks that cannot be solved by simple machine algorithms (Li et al. 2016). A large body of work exists around the integration of human inputs into microtask crowdsourcing environments (Lasecki 2014). Consistently, many studies attempt to tackle tasks that can be easily decomposed into simpler subtasks and accomplished independently (Cheng et al. 2015). With the growth of expert crowdsourcing settings comprising non-decomposable macrotasks, there is an increasing need to support complex work (Schmitz and Lykourantzou 2018). Such open-ended worker inputs often implicate a high level of dependency and expertise (Zakaria and

A. Correia (✉) · H. Paredes · B. Fonseca
University of Trás-os-Montes e Alto Douro, UTAD, Quinta de Prados, Apartado 1013, Vila Real,
Portugal
e-mail: antonio.g.correia@inesctec.pt

INESC TEC, Porto, Portugal

A. Correia · S. Jameel
University of Kent, School of Computing, Medway Campus, Canterbury, UK

D. Schneider
Tércio Pacitti Institute of Computer Applications and Research (NCE), Federal University of Rio
de Janeiro, Rio de Janeiro, Brazil

© Springer Nature Switzerland AG 2019

V.-J. Khan et al. (eds.), *Macrotask Crowdsourcing*,

Human-Computer Interaction Series, https://doi.org/10.1007/978-3-030-12334-5_5

Abdullah 2018). In particular, macrotasking projects go beyond data processing to produce new information through social interaction among crowd members (Walsh et al. 2014). The crowdsourcing tasks in this line of work need contextual information and can imply overheads regarding the increasing levels of coordination required to generate information socially. Haas and colleagues (2015) go even further by arguing that “a key challenge in macrotask-powered work is evaluating the quality of a worker’s output” due to the absence of an aggregation method in which the inputs of the crowd can be easily combined and evaluated. In this sense, we should state at the outset that supporting macrotasks is particularly challenging and there is still a need to identify new pathways along which complex crowd work can be effectively accomplished.

This work furthers an existing strand of research in leveraging collaborative efforts between humans and machine agents to handle the complexity of the work that can be performed by IT-mediated crowds in science. Crowdsourcing has been successfully used as a tool for supporting scientific research (Law et al. 2017), and research problems of massive scale can be distributed among a sizeable pool of experts and volunteers who contribute actively by handling massive quantities of assorted data (Hochachka et al. 2012). Researchers attempting to perform complex scientific tasks (e.g., systematic literature reviews) usually decompose them into smaller, more manageable chunks of work that can be used to generate training data for AI algorithms (Krivosheev et al. 2018). Such small-scale scientific work settings must be further expanded to incorporate the untapped potentials of combining crowd interactions with automated reasoning at a large-scale given the value of the crowd-AI integration to produce large amounts of data and attain novel discoveries on multivariate topics. Adding on to this line of inquiry, this chapter explores some theoretical underpinnings of crowd-AI hybrids in the context of complex work while depicting a research agenda with a vast set of gaps reported in the literature.

The rest of this chapter proceeds as follows. In Sect. 5.2 we present some background on macrotask crowdsourcing and hybrid machine-crowd interaction in the context of scientific work. In this section, we also illustrate examples of current systems and frameworks intended to support expert crowdsourcing. In Sect. 5.3, we describe some design claims and general aspects of crowdsourcing and AI applications found in the literature. We close in Sect. 5.4 with some remarks and future directions on the combination of crowd-computing hybrids.

5.2 Macrotask Crowdsourcing in Science: From HCI to Hybrid Crowd-Machine Applications

As the number of publications continues to increase, discovery, and acquisition of useful scholarly data from academic literature impose several challenges (Dong et al. 2017). In addition, a large amount of resources are usually spent on research practices (Rigby 2009). Crowd science can be a trustworthy solution for tackling scientific

problems that are beyond the capabilities of computer algorithms by engaging academic researchers and nonprofessional scientists (Franzoni and Sauermann 2014). Although several research studies have demonstrated the potential uses of crowdsourcing in science, many researchers are still reluctant regarding the adoption of crowdsourcing (Law et al. 2017). Researchers have been studying crowdsourcing as a way to reduce the cost and speed of a research project while enhancing the quality of the work (Ranard et al. 2014; Tsueng et al. 2016). On the other hand, reviews of the prior research on crowdsourcing show that there are some challenges on scaling up complexity maintaining high-quality responses (Barowy et al. 2012). Human-Computer Interaction (HCI) reaching a high level of engagement over time is another concern in crowd science (Nov et al. 2014). Past research in HCI has explored the use of platforms like Amazon Mechanical Turk (AMT)¹ for crowdsourcing research. For example, Good et al. (2014) recruited nonscientists to identify disease concepts in biomedical paper abstracts and showed that crowd-powered systems can be a reliable instrument for creating annotated corpora. Basing their approach on the general assumption that crowd annotations can be of equal (or even better) value when compared to experts, several authors have used AMT to systematically evaluate scientific literature (e.g., Brown and Allison 2014; Mortensen et al. 2017; Krivosheev et al. 2017). Nevertheless, very little is known about the adoption of alternative platforms such as Prolific Academic² and Crowdcrafting³ for crowdsourcing research (Peer et al. 2017). While this is an obvious limitation, there are several reasons why this fact may be acceptable. In comparison to other crowdsourcing platforms used for research, these platforms usually lack a large and active user base and a suitable API to programmatically access the platform's functionalities.

As previously noted, crowdsourcing tasks can be categorized into microtasks and macrotasks (Luz et al. 2015). Microtask-level settings are characterized by repetitive tasks that are simple for individuals to perform (e.g., image labeling). Such tasks comprise context-free units of work, do not require special skills, and the reward for each task is usually small (Xie and Lui 2018). In macrotasking, requesters create high-level tasks without microtask decomposition while paying workers fair hourly wages (Marcus and Parameswaran 2015). In the literature, there are several examples of expert crowdsourcing systems and general online macrotask-powered work platforms (see Table 5.1). As the table shows, these tools differ from microtasking platforms due to their focus on solving innovative and complex tasks that require high levels of expertise to complete. In contrast to AMT, expert crowdsourcing platforms allow requesters and workers to participate in persistent one-on-one discussions (Salehi et al. 2017). The macrotasks supported by these platforms are usually freeform and large in the sense that they need a vast amount of time to complete.

Macrotasks have particular dependencies, changing requirements, and require expert skills and varied types of expertise. In addition, they are socially mediated in the sense that they require collaboration and may take more time to complete

¹<https://www.mturk.com/>.

²<https://prolific.ac/>.

³<https://crowdcrafting.org/>.

Table 5.1 Examples of crowd-powered systems and frameworks for supporting macrotasks

	System	Description	Reference	
Global online macrotask powered work platforms and commercial products	Fiverr	<i>A platform for outsourcing challenging and innovative tasks</i>	Xie & Lui (2018)	
	Upwork	<i>Generic online outsourcing marketplace for creative tasks</i>	Marcus & Parameswaran (2015)	
	OpenIDEO	<i>Social innovation platform for collaboratively tackling global issues</i>	Schmitz & Lykourantzou (2016)	
	Freelancer	<i>Global online work platform for freelancers</i>	Borromeo & Toyama (2016)	
	Quirky	<i>Community-led invention platform for product design</i>	Schmitz & Lykourantzou (2016)	
	Science Exchange	<i>Outsourcing platform for solving scientific problems through a network of qualified crowd workers</i>	Yan et al. (2016)	
	Crowdspring	<i>Online marketplace for crowdsourced creative designs and ideas</i>	Schmitz & Lykourantzou (2016)	
	Innocentive	<i>Open innovation marketplace for tackling complex problems</i>	Steg et al. (2010)	
	Argonaut	<i>Framework that uses hierarchical review to improve complex work</i>	Haas et al. (2015)	
	CrowdWeaver	<i>A system to track crowd workers and task progress</i>	Kittur et al. (2012)	
	Crowd4U	<i>A prototype system for the deployment of collaborative tasks</i>	Ikeda et al. (2016)	
	CrowdSCIM	<i>Allows novice crowd workers to learn historical thinking skills while completing historical research tasks</i>	Wang et al. (2018)	
	Crowdforge	<i>A framework for performing complex work and interdependent tasks</i>	Kittur et al. (2011)	
	Expert crowdsourcing systems	Data Tamer	<i>Expert crowdsourcing system for handling uncertainties in entity resolution and schema integration</i>	Stonebraker et al. (2013)
Prism		<i>Allows a user to upload text documents for collective interpretation</i>	Walsh et al. (2014)	
MobileWorks		<i>Operates as an algorithmically managed service, routing work to qualified participants</i>	Kulkarni et al. (2012)	
Crowd		<i>A platform for crowdsourcing complex work where one can submit a workflow</i>	Chettih et al. (2014)	
WearWrite		<i>Allows a user to perform complex tasks (e.g., writing) from a wearable device</i>	Nebeling et al. (2015)	
Wish		<i>A system that uses expert crowd members to carry out complex (creative) tasks</i>	Kulkarni et al. (2014)	
Crowd-AI hybrid systems in science		PANDA	<i>Hybrid, crowd-computing system for academic knowledge discovery and acquisition</i>	Dong et al. (2017)
		Solvent	<i>Mixed-initiative system for finding analogies between research papers</i>	Chan et al. (2018)
		CrowdRev	<i>A platform for crowd-enabled screening of systematic literature reviews</i>	Ramirez et al. (2018)
		SciCrowd	<i>Crowd-AI system for supporting research work in academic settings</i>	Correia et al. (2018a)
	Apollo	<i>A mixed-initiative system that interactively explores large networks of scientific data</i>	Chou et al. (2011)	

(Schmitz and Lykourantzou 2018). In macrotasking settings, requesters reward workers according to the quality of the evaluated solution. In this sense, a requester will only give a large reward to a worker if the quality of the solution is high. Workflows are needed to facilitate the decomposition of tasks into subtasks, management of dependencies between subtasks, and assembly of results (Kittur et al. 2013). Current approaches do not encompass human factors in assessing the quality of the solution, do not address the challenge of free riding of workers, nor denial of payment of requesters (Xie and Lui 2018). Macrotask crowdsourcing for complex work cannot be realized by using simple parallel approaches like aggregating multiple independent judgments through voting since macrotasks are difficult to be decomposed and require sharing of contextual information. As argued by Niu et al. (2018), a crowd may need to build its own team for solving complex tasks.

Research is beginning to emerge in exploring ways to optimize macrotasking scenarios. Retelny and co-workers (2014) proposed *flash teams*, a framework that relies on expert crowdsourcing for solving tasks that require deep domain knowledge. Recently, Valentine and colleagues (2017) extended this expert crowd work framework to *flash organizations*, an approach where crowds are “structured like organizations to achieve complex and open-ended goals”. CrowdForge (Kittur et al. 2011) is another example of a framework for executing complex tasks that incorporates some level of automation in the coordination process (Garcia-Molina et al. 2016). In the same vein, Prism (Walsh et al. 2014) was introduced as a system with a shared digital space in which crowd workers can provide creative contributions and interpretations of texts. Argonaut (Haas et al. 2015) is perhaps one of the most widely known examples of a macrotask crowdsourcing system introduced in the literature. The system is intended to support context-aware data processing tasks through a hierarchical review scheme. Platforms such as Crowd (Chettih et al. 2014), Wish (Kulkarni et al. 2014), MobileWorks (Kulkarni et al. 2012), and Data Tamer (Stonebraker et al. 2013) also represent the vast range of solutions that leverage a crowd of domain experts to carry out macrotasks.

The ongoing stream of publications about macrotasking also suggests the use of such applications for learning and research purposes. Crowd4U (Morishima et al. 2012) is a complex data-centric crowdsourcing system that supports collaborative tasks by enabling task decomposition and assignment. Furthermore, CrowdSCIM (Wang et al. 2018) enables a vast set of macrotasks to improve historical research tasks without feedback or intervention from other crowd members. To achieve the full potential of crowdsourcing in science, HCI researchers have also shown a variety of scenarios in which crowd members can be engaged in advanced research tasks such as writing a paper (Gaikwad et al. 2016; Whiting et al. 2017; Crowston et al. 2018). There are other examples of hybrid crowd-AI systems proposed for supporting complex scientific tasks, as can be seen in Table 5.1. To tackle the problem of academic knowledge acquisition, PANDA (Dong et al. 2017) combines hybrid algorithmic-crowdsourcing techniques, while SciCrowd (Correia et al. 2018a) supports research groups on data-driven research tasks (e.g., annotation of large amounts of HCI publications) taking into account a particular research question instead of a simple search for terms. Concomitantly, in research, we have seen systems where

humans can annotate aspects of research papers (e.g., findings) in order to find analogies through a computational model (Chan et al. 2018). Others in the community have studied how to combine machine and crowd intelligence in systematic literature reviews (Ramirez et al. 2018). At the same time, Nguyen et al. (2015) combined active learning, domain experts, and crowd workers to support citation screening in systematic literature reviews. However, many aspects regarding crowd-AI interaction have not been investigated by the HCI community intensively so far. While several papers touch on issues of algorithmic crowd-AI hybrids, supporting research macro-tasks was not the focus of existing literature since it has predominantly discussed the technology driving mechanisms in microtasking scenarios with little detail on how technology has been adopted as well as the socio-technical aspects required to facilitate a crowd-AI integration for solving complex problems in science.

5.3 Crowd-AI Systems as a Scaffold for Complex Work

When applied to highly complex problem-solving tasks, the depth and breadth of crowd-powered systems are far beyond the traditional definition of macrotask crowdsourcing. In some circumstances, they can benefit from a crowd-AI hybrid approach. However, replicating one second of human brain activity corresponds to more than 80,000 processors and over a petabyte of system memory (Gil and Hirsh 2012). This involves a vast set of challenges for deploying AI algorithms able to systematically explore multidimensional data and autonomously discover patterns at large scale (Gil et al. 2014). On reading the literature, a significant body of research exists on the adoption of crowd intelligence as a scaffold for machine learning (Kamar 2016). For instance, crowd-machine systems like Flock (Cheng and Bernstein 2015) combine the strengths of human crowd workers and computer algorithms to generate hybrid classifiers. As shown in Table 5.2, there are also some design issues that can be taken into account in the deployment of macrotask crowdsourcing systems.

With the rapid growth of crowdsourcing, many scholars have exhaustively discussed aspects such as crowdsourced task features, quality control, crowd and crowd-sourcer attributes, motivational factors, crowdsourcing system features, role of contributors, and aggregation mechanisms, to name a few (e.g., Vukovic 2009; Geiger et al. 2011; Dong et al. 2017). Crowd workers can collaborate explicitly to solve a target problem by sharing structured information or building artifacts (e.g., software). On the other hand, implicit collaboration involves “invisible” contributions such as solving *captchas* and play games with a scientific purpose (Doan et al. 2011). A task can vary in terms of complexity (e.g., routine), variety, modularity, solvability (e.g., simple to humans), structure, and reliability (Hosseini et al. 2014). A task may be also difficult or expensive to automate. Task dependency represents a critical aspect of macrotasks since crowd workers need to coordinate and build upon the contributions of the other members (Schmitz and Lykourantzou 2016). Some macrotasks (e.g., perform a qualitative study in the field of HCI) are not easily decomposable (Krivosheev et al. 2018) and a critical factor in crowdsourcing complex work relies

Table 5.2 Design framework of macrotask crowdsourcing systems

<i>Dimensions of macrotask crowdsourcing systems</i>	<i>Characteristics of crowd-computing applications</i>
Type of target problem	Application class
Crowdsourcing type	Predetermined purpose
> Passive, Directive, Collaborative	Ubiquity
Nature of collaboration	Persuasion
> Implicit, Explicit	Cloud computing
Communication	Collective intelligence
Coordination	Interests sharing
Modularity	Latency
> Decomposable, Non-decomposable	Integrating human inputs into AI systems
Complexity	> Reasoning abilities for hybrid intelligence, Training
> Routine, Complex, Creative	User interface
Degree of manual effort	Architecture
Recruit and retain users	Connectivity
Combine/aggregate inputs	-----
Quality control	<i>Crowdware time-space matrix</i>
Crowd size and roles	-----
Expertise	Same place, Different places
Feedback	> Virtual, Physical
-----	Same time, Different times
<i>Implications for designing crowd-powered systems</i>	> Critical, Non-critical
-----	-----
Timing	<i>Crowdsourcing platform features (facilities)</i>
Scalability	-----
Locality	Computing platform
Reliability	> Internal, External
Synergy	Type of platform
Dependency	Distinguishing features/facilities
Automation	> Crowd-related interactions
> Simple, Difficult	> Crowdsourcer-related interactions
Transparency	> Task-related facilities
Crowd work regulation and ethics	> Platform-related facilities
Anonymity and privacy	System and technology issues
-----	Common design patterns
<i>Design patterns embodied in programming metaphors</i>	Openness
-----	> Closed, Internally open, Externally open
Idea ecology	Ownership
Web of dependencies	> Public, Private
Intellectual supply chain	Highlighting data use
Collaborative deliberation	Technology access and proficiency of potential participants
Radically fluid virtual organization	Accessible crowd work and assistive technology
Multi-user games	-----

on the ability of coordinating crowds by means of reliable tasks, protocols, and feedback (Vaish et al. 2017). As argued by Weiss (2016), crowdsourcing approaches also differ in terms of the type of tasks assigned to the crowd, the amount of time spent, and the level of collaboration between members.

The behavior of a crowd in a crowdsourcing system can also vary taking into account its architecture (Doan et al. 2011). For example, a standalone system deals with challenges like recruiting participants and choosing their potential actions. As a large group of individuals with a shared purpose and emotions, a crowd can be physically or virtually situated and the nature of the task is an influential factor concerning the way in which crowd members might be engaged (Schneider et al. 2012). Previous research has also suggested that crowd workers are classified in terms of diversity, largeness, unknownness, underfinedness, and suitability (Hosseini et al. 2014). In crowdsourcing research settings, possible roles include principal researcher, research assistant, and participants or members of the crowd (Vaish et al. 2017) who have different abilities (e.g., pattern recognition) and use computing devices to interact, coordinate and execute tasks (Parshotam 2013). According to Bigham et al. (2015), there are three main types of crowdsourcing. In passive crowdsourcing, crowd participants are unknown to each other but there is the possibility of tracing their collective behavior. Directed crowdsourcing relies on the recruitment and guidance of crowd

members through a single individual or algorithm. In collaborative crowdsourcing, the coordination tasks are usually performed by a group of individuals with a shared purpose and a self-determined structure (e.g., Wikipedia⁴).

Concerning the characteristics of crowd-computing applications, scalability is a key feature for crowd-AI hybrids in the sense that we need to adapt to different situations and levels of complexity (Talia 2019). Scaling up the crowd reduces the downtime and thus decreases the latency in crowdsourcing (Difallah et al. 2014). The machine must also provide feedback to the user by interactively informing the decision-making process. In a hybrid crowd-AI system such as CrowdFlow (Quinn et al. 2010), complex crowd work outputs are used to provide feedback for machine algorithms and thus enhance their algorithmic power. Dow and colleagues (2012) identified key dimensions of crowd feedback, including timeliness (asynchronous, synchronous), specificity, source (e.g., peer workers), and format. Prior research also suggests that social transparency among crowd workers can be particularly beneficial in crowdsourcing settings (Huang and Fu 2013). Nonetheless, such mechanisms must be implemented with caution to prevent malicious behaviors in crowd-AI interaction (Kittur et al. 2013).

A large body of work (e.g., Hetmank 2013; Daniel et al. 2018) has exploited the use of new techniques for aggregating crowd inputs while controlling the quality of the contributions and the reliability of contributors as critical factors to the success of crowdsourcing since the responses provided by crowd members can be error-prone and biased due to malicious (or less motivated) workers (Lasecki et al. 2014). This calls into question a number of assumptions that lie behind the notion of “quality control”. Daniel et al.’s (2018) investigation on quality attributes and assessment techniques found that quality assessment methods range from self-assessment to peer review, voting, gold standards, and feedback aggregation. Crowd participants are usually engaged in complex work through intrinsic motivational factors (e.g., passion, enjoyment and fun, sense of community, personal achievement) and extrinsic motivations such as financial rewards and promotion (Geiger et al. 2011). In addition, crowd work regulation and ethics raise a lot of concerns about privacy and anonymity, worker rights and fair wages, discrimination, and intellectual property (Hansson and Ludwig 2018). In this kind of scenario, sensitive information about crowd workers such as home location and hobbies can be retrieved and used improperly. Furthermore, we should state at the outset that accessible crowd work (Zyskowski et al. 2015) must be leveraged by assistive technology to support people with disabilities and special needs.

An earlier review of the literature on the design components of crowdsourcing platforms (Hetmank 2013) revealed a focus on the functions and operations of a crowd-powered system as an intermediary that distributes Human Intelligence Tasks (HITs) from requesters to the crowd workers. A crowdsourcing system also comprises technical attributes such as software components, functions, and data objects. As these technologies develop, attention to the design processes that support their outputs is essential. Developers of crowd-powered systems must pay attention to

⁴<https://www.wikipedia.org/>

aspects like awareness, user interface, authentication, quality control, and workflow support. Typically, workflow systems are deployed “ad hoc” and tailored to particular use cases (Lofi and El Maarry 2014). A crowdsourcing platform must support actions such as recruit and evaluate crowd workers, define and assign HITs, submit contributions, set time period, state rewards, and pay crowd workers. As argued by Vukovic (2009), the loss of network connectivity can compromise the interaction in real-time crowdsourcing settings where a failure may be critical to human lives, as in the case of crisis and emergency response.

By virtue of the recent research efforts on crowd-AI hybrids, there are several missing pieces and areas for future work. The literature on this topic is limited and great care must be taken to aspects like task design (Vaish et al. 2015), risk of overspecialization and failing heuristics (Lofi and El Maarry 2014), ambiguity and systematic error biases (Vaughan 2018), and overload of crowd-generated inputs (Barbier et al. 2012). Some requirements for crowd-AI systems include the translation of system states and operations between humans and machines by means of contextual information (Dong et al. 2017) and the adequate support for open-ended, complex scientific activities at different scales (Correia et al. 2018b). These concerns are often overlooked and result from the increasing complexity of algorithms. Within HCI, the adoption of interactive, human-guided machine learning (Gil et al. 2019) constitutes further avenues of research into the intersection of crowdsourcing and AI for supporting macrotasks.

5.4 Final Remarks

In this chapter, we addressed the need for handling complexity in crowd work through the integration of crowd-AI hybrids. This approach appears to be a viable solution for many areas. Nonetheless, we are aware of very little work that tries to characterize such kind of combination in the context of macrotask crowdsourcing as it moves on from its young age. In framing it as a problem, we want to explore the ways in which the design of intelligent systems can be informed by symbiotic interactions between crowds and machines able to completing complex tasks. The full extent of this crowd-guided AI model will be studied in future stages of this research towards a conceptual framework predicated on the socio-technical aspects that need to be considered when solving complex tasks that require high levels of interdependency and domain expertise.

References

- Barbier, G., Zafarani, R., Gao, H., Fung, G., & Liu, H. (2012). Maximizing benefits from crowd-sourced data. *Computational and Mathematical Organization Theory*, 18(3), 257–279.

- Barowy, D. W., Curtsinger, C., Berger, E. D., & McGregor, A. (2012). Automan: A platform for integrating human-based and digital computation. *ACM SIGPLAN Notices*, 47(10), 639–654.
- Bigam, J. P., Bernstein, M. S., & Adar, E. (2015). Human-computer interaction and collective intelligence. *Handbook of Collective Intelligence*, 57.
- Borromeo, R. M., & Toyama, M. (2016). An investigation of unpaid crowdsourcing. *Human-Centric Computing and Information Sciences*, 6(1), 11.
- Brown, A. W., & Allison, D. B. (2014). Using crowdsourcing to evaluate published scientific literature: Methods and example. *PLoS ONE*, 9(7), e100647.
- Chan, J., Chang, J. C., Hope, T., Shahaf, D., & Kittur, A. (2018). Solvent: A mixed initiative system for finding analogies between research papers. In *Proceedings of the ACM Conference on Computer-Supported Cooperative Work and Social Computing*.
- Chau, D. H., Kittur, A., Hong, J. I., & Faloutsos, C. (2011). Apolo: Making sense of large network data by combining rich user interaction and machine learning. In *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems* (pp. 167–176).
- Cheng, J., & Bernstein, M. S. (2015). Flock: Hybrid crowd-machine learning classifiers. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work and Social Computing* (pp. 600–611).
- Cheng, J., Teevan, J., Iqbal, S. T., & Bernstein, M. S. (2015). Break it down: A comparison of macro-and microtasks. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (pp. 4061–4064).
- Chettih, A., Gross-Amblard, D., Guyon, D., Legeay, E., & Miklós, Z. (2014). Crowd, a platform for the crowdsourcing of complex tasks. In *BDA 2014: Gestion de Données—Principes, Technologies et Applications* (pp. 51–55).
- Correia, A., Schneider, D., Paredes, H., & Fonseca, B. (2018a). SciCrowd: Towards a hybrid, crowd-computing system for supporting research groups in academic settings. In *Proceedings of the 24th International Conference on Collaboration and Technology* (pp. 34–41).
- Correia, A., Schneider, D., Fonseca, B., & Paredes, H. (2018b). Crowdsourcing and massively collaborative science: A systematic literature review and mapping study. In *Proceedings of the 24th International Conference on Collaboration and Technology* (pp. 133–154).
- Crowston, K., Mitchell, E., & Østerlund, C. (2018). Coordinating advanced crowd work: Extending citizen science. In *Proceedings of the 51st Hawaii International Conference on System Sciences* (pp. 1681–1690).
- Daniel, F., Kucherbaev, P., Cappiello, C., Benatallah, B., & Allahbakhsh, M. (2018). Quality control in crowdsourcing: A survey of quality attributes, assessment techniques, and assurance actions. *ACM Computing Surveys (CSUR)*, 51(1), 7.
- Difallah, D. E., Catasta, M., Demartini, G., & Cudré-Mauroux, P. (2014). Scaling-up the crowd: Micro-task pricing schemes for worker retention and latency improvement. In *Second AAAI Conference on Human Computation and Crowdsourcing*.
- Doan, A., Ramakrishnan, R., & Halevy, A. Y. (2011). Crowdsourcing systems on the world-wide web. *Communications of the ACM*, 54(4), 86–96.
- Dong, Z., Lu, J., Ling, T. W., Fan, J., & Chen, Y. (2017). Using hybrid algorithmic-crowdsourcing methods for academic knowledge acquisition. *Cluster Computing*, 20(4), 3629–3641.
- Dow, S., Kulkarni, A., Klemmer, S., & Hartmann, B. (2012). Shepherding the crowd yields better work. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work* (pp. 1013–1022).
- Franzoni, C., & Sauermann, H. (2014). Crowd science: The organization of scientific research in open collaborative projects. *Research Policy*, 43(1), 1–20.
- Gaikwad, S. N. S., Morina, D., Ginzberg, A., Mullings, C., Goyal, S., Gamage, D., et al. (2016). Boomerang: Rebounding the consequences of reputation feedback on crowdsourcing platforms. In *Proceedings of the 29th ACM Symposium on User Interface Software and Technology* (pp. 625–637).
- Garcia-Molina, H., Joglekar, M., Marcus, A., Parameswaran, A., & Verroios, V. (2016). Challenges in data crowdsourcing. *IEEE Transactions on Knowledge and Data Engineering*, 28(4), 901–911.

- Geiger, D., Seedorf, S., Schulze, T., Nickerson, R. C., & Schader, M. (2011). Managing the crowd: Towards a taxonomy of crowdsourcing processes. In *Proceedings of the Proceedings of the 17th Americas Conference on Information Systems*.
- Gil, Y., & Hirsh, H. (2012). Discovery informatics: AI opportunities in scientific discovery. In *Proceedings of the AAAI Fall Symposium: Discovery Informatics*.
- Gil, Y., Greaves, M., Hendler, J., & Hirsh, H. (2014). Amplify scientific discovery with artificial intelligence. *Science*, 346(6206), 171–172.
- Gil, Y., Honaker, J., Gupta, S., Ma, Y., D’Orazio, V., Garijo, D., et al. (2019). Towards human-guided machine learning. In *Proceedings of the 24th ACM International Conference on Intelligent User Interfaces*.
- Good, B. M., Nanis, M., Wu, C., & Su, A. I. (2014). Microtask crowdsourcing for disease mention annotation in PubMed abstracts. In *Proceedings of the Pacific Symposium on Biocomputing* (pp. 282–293).
- Haas, D., Ansel, J., Gu, L., & Marcus, A. (2015). Argonaut: Macrotask crowdsourcing for complex data processing. *Proceedings of the VLDB Endowment*, 8(12), 1642–1653.
- Hansson, K., & Ludwig, T. (2018). Crowd dynamics: Conflicts, contradictions, and community in crowdsourcing. *Computer Supported Cooperative Work (CSCW)*, 1–4.
- Hetmank, L. (2013). Components and functions of crowdsourcing systems – A systematic literature review. *Wirtschaftsinformatik*, 4.
- Hochachka, W. M., Fink, D., Hutchinson, R. A., Sheldon, D., Wong, W. K., & Kelling, S. (2012). Data-intensive science applied to broad-scale citizen science. *Trends in Ecology & Evolution*, 27(2), 130–137.
- Hosseini, M., Phalp, K., Taylor, J., & Ali, R. (2014). The four pillars of crowdsourcing: A reference model. In *Proceedings of the 2014 IEEE Eighth International Conference on Research Challenges in Information Science (RCIS)* (pp. 1–12).
- Huang, S. W., & Fu, W. T. (2013). Don’t hide in the crowd!: Increasing social transparency between peer workers improves crowdsourcing outcomes. In *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems* (pp. 621–630).
- Ikeda, K., Morishima, A., Rahman, H., Roy, S. B., Thirumuruganathan, S., Amer-Yahia, S., et al. (2016). Collaborative crowdsourcing with Crowd4U. *Proceedings of the VLDB Endowment*, 9(13), 1497–1500.
- Kamar, E. (2016). Directions in hybrid intelligence: Complementing AI systems with human intelligence. In *IJCAI* (pp. 4070–4073).
- Kittur, A., Smus, B., Khamkar, S., & Kraut, R. E. (2011). Crowdforge: Crowdsourcing complex work. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology* (pp. 43–52).
- Kittur, A., Khamkar, S., André, P., & Kraut, R. (2012). CrowdWeaver: Visually managing complex crowd work. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work* (pp. 1033–1036).
- Kittur, A., Nickerson, J. V., Bernstein, M., Gerber, E., Shaw, A., Zimmerman, J., et al. (2013). The future of crowd work. In *Proceedings of the ACM Conference on Computer-Supported Cooperative Work and Social Computing* (pp. 1301–1318).
- Krivosheev, E., Casati, F., Caforio, V., & Benatallah, B. (2017). *Crowdsourcing paper screening in systematic literature reviews*. [arXiv:1709.05168](https://arxiv.org/abs/1709.05168).
- Krivosheev, E., Casati, F., & Benatallah, B. (2018). Crowd-based multi-predicate screening of papers in literature reviews. In *Proceedings of the World Wide Web Conference* (pp. 55–64).
- Kulkarni, A., Gutheim, P., Narula, P., Rolnitzky, D., Parikh, T., & Hartmann, B. (2012). Mobile-works: Designing for quality in a managed crowdsourcing architecture. *IEEE Internet Computing*, 16(5), 28–35.
- Kulkarni, A., Narula, P., Rolnitzky, D., & Kontny, N. (2014). Wish: Amplifying creative ability with expert crowds. In: *Second AAAI Conference on Human Computation and Crowdsourcing*.
- Lasecki, W. S. (2014). Crowd-powered intelligent systems. *Human Computation Journal*.

- Lasecki, W. S., Teevan, J., & Kamar, E. (2014). Information extraction and manipulation threats in crowd-powered systems. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work and Social Computing* (pp. 248–256).
- Law, E., Gajos, K. Z., Wiggins, A., Gray, M. L., & Williams, A. C. (2017). Crowdsourcing as a tool for research: Implications of uncertainty. In *Proceedings of the ACM Conference on Computer-Supported Cooperative Work and Social Computing* (pp. 1544–1561).
- Li, G., Wang, J., Zheng, Y., & Franklin, M. J. (2016). Crowdsourced data management: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 28(9), 2296–2319.
- Lofi, C., & El Maarry, K. (2014). Design patterns for hybrid algorithmic-crowdsourcing workflows. *CBI*, 1 (pp. 1–8).
- Luz, N., Silva, N., & Novais, P. (2015). A survey of task-oriented crowdsourcing. *Artificial Intelligence Review*, 44(2), 187–213.
- Marcus, A., & Parameswaran, A. (2015). Crowdsourced data management: Industry and academic perspectives. *Foundations and Trends in Databases*, 6(1–2), 1–161.
- Morishima, A., Shinagawa, N., Mitsuishi, T., Aoki, H., & Fukusumi, S. (2012). CyLog/Crowd4U: A declarative platform for complex data-centric crowdsourcing. *Proceedings of the VLDB Endowment*, 5(12), 1918–1921.
- Mortensen, M. L., Adam, G. P., Trikalinos, T. A., Kraska, T., & Wallace, B. C. (2017). An exploration of crowdsourcing citation screening for systematic reviews. *Research Synthesis Methods*, 8(3), 366–386.
- Nebeling, M., Guo, A., To, A., Dow, S., Teevan, J., & Bigham, J. (2015). WearWrite: Orchestrating the crowd to complete complex tasks from wearables. In *Adjunct Proceedings of the 28th Annual ACM Symposium on User Interface Software and Technology* (pp. 39–40).
- Nguyen, A. T., Wallace, B. C., & Lease, M. (2015). Combining crowd and expert labels using decision theoretic active learning. In *Proceedings of the Third AAAI Conference on Human Computation and Crowdsourcing*.
- Niu, X. J., Qin, S. F., Vines, J., Wong, R., & Lu, H. (2018). Key crowdsourcing technologies for product design and development. *International Journal of Automation and Computing*, 1–15.
- Nov, O., Arazy, O., & Anderson, D. (2014). Scientists@Home: What drives the quantity and quality of online citizen science participation? *PLoS ONE*, 9(4), e90375.
- Parshotam, K. (2013). Crowd computing: A literature review and definition. In *Proceedings of the South African Institute for Computer Scientists and Information Technologists Conference* (pp. 121–130).
- Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017). Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, 70, 153–163.
- Quinn, A. J., Bederson, B. B., Yeh, T., & Lin, J. (2010). CrowdfLOW: Integrating machine learning with Mechanical Turk for speed-cost-quality flexibility. *Better Performance over Iterations*.
- Ramirez, J., Krivosheev, E., Baez, M., Casati, F., & Benatallah, B. (2018). CrowdRev: A platform for crowd-based screening of literature reviews. [arXiv:1805.12376](https://arxiv.org/abs/1805.12376).
- Ranard, B. L., Ha, Y. P., Meisel, Z. F., Asch, D. A., Hill, S. S., Becker, L. B., et al. (2014). Crowdsourcing—Harnessing the masses to advance health and medicine, a systematic review. *Journal of General Internal Medicine*, 29(1), 187–203.
- Retelny, D., Robaszekiewicz, S., To, A., Lasecki, W. S., Patel, J., Rahmati, N., & Bernstein, M. S. (2014). Expert crowdsourcing with flash teams. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology* (pp. 75–85).
- Rigby, J. (2009). Comparing the scientific quality achieved by funding instruments for single grant holders and for collaborative networks within a research system: Some observations. *Scientometrics*, 78(1), 145–164.
- Salehi, N., Teevan, J., Iqbal, S., & Kamar, E. (2017). Communicating context to the crowd for complex writing tasks. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work and Social Computing* (pp. 1890–1901).
- Schmitz, H., & Lykourantzou, I. (2016). *It's about time: Online macrotask sequencing in expert crowdsourcing*. [arXiv:1601.04038](https://arxiv.org/abs/1601.04038).

- Schmitz, H., & Lykourantzou, I. (2018). Online sequencing of non-decomposable macrotasks in expert crowdsourcing. *ACM Transactions on Social Computing*, 1(1), 1.
- Schneider, D., Moraes, K., De Souza, J. M., & Esteves, M. G. P. (2012). CSCWD: Five characters in search of crowds. In *Proceedings of the IEEE International Conference on Computer Supported Cooperative Work in Design* (pp. 634–641).
- Sieg, J. H., Wallin, M. W., & von Krogh, G. (2010). Managerial challenges in open innovation: A study of innovation intermediation in the chemical industry. *R&D Management*, 40(3), 281–291.
- Stonebraker, M., Bruckner, D., Ilyas, I. F., Beskales, G., Cherniack, M., Zdonik, S. B. et al. (2013). Data curation at scale: The data tamer system. In *CIDR*.
- Talia, D. (2019). A view of programming scalable data analysis: From clouds to exascale. *Journal of Cloud Computing*, 8(1), 4.
- Tsueng, G., Nanis, M., Fouquier, J., Good, B., & Su, A. (2016). Citizen science for mining the biomedical literature. *BioRxiv*, 038083.
- Vaish, R., Davis, J., & Bernstein, M. (2015). Crowdsourcing the research process. *Collective Intelligence*.
- Vaish, R., Gaikwad, S. N. S., Kovacs, G., Veit, A., Krishna, R., Arrieta Ibarra, I.,... & Davis, J. (2017). Crowd research: Open and scalable university laboratories. In: *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology* (pp. 829–843).
- Valentine, M. A., Retelny, D., To, A., Rahmati, N., Doshi, T., & Bernstein, M. S. (2017). Flash organizations: Crowdsourcing complex work by structuring crowds as organizations. In *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems* (pp. 3523–3537).
- Vaughan, J. W. (2018). Making better use of the crowd: How crowdsourcing can advance machine learning research. *Journal of Machine Learning Research*, 18(193), 1–46.
- Vukovic, M. (2009). Crowdsourcing for enterprises. In *IEEE Congress on Services-I* (pp. 686–692).
- Walsh, B., Maiers, C., Nally, G., Boggs, J., & Team, Praxis Program. (2014). Crowdsourcing individual interpretations: Between microtasking and macrotasking. *Literary and Linguistic Computing*, 29(3), 379–386.
- Wang, N. C., Hicks, D., & Luther, K. (2018). Exploring trade-offs between learning and productivity in crowdsourced history. In *Proceedings of the ACM on Human-Computer Interaction (CSCW)* (Vol. 2, p. 178).
- Weiss, M. (2016). Crowdsourcing literature reviews in new domains. *Technology Innovation Management Review*, 6(2), 5–14.
- Whiting, M. E., Gamage, D., Gaikwad, S. N. S., Gilbee, A., Goyal, S., Ballav, A., et al. (2017). Crowd guilds: Worker-led reputation and feedback on crowdsourcing platforms. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work and Social Computing* (pp. 1902–1913).
- Xie, H., & Lui, J. C. (2018). Incentive mechanism and rating system design for crowdsourcing systems: Analysis, tradeoffs and inference. *IEEE Transactions on Services Computing*, 11(1), 90–102.
- Yan, X., Ding, X., & Gu, N. (2016). Crowd work with or without crowdsourcing platforms. In: *Proceedings of the IEEE 20th International Conference on Computer Supported Cooperative Work in Design (CSCWD)* (pp. 56–61).
- Zakaria, N. A., & Abdullh, C. Z. H. (2018). Crowdsourcing and library performance in digital age. *Development*, 7(3).
- Zyskowski, K., Morris, M. R., Bigham, J. P., Gray, M. L., & Kane, S. K. (2015). Accessible crowdwork? Understanding the value in and challenge of microtask employment for people with disabilities. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work and Social Computing* (pp. 1682–1693).

Chapter 6

What You Sow, So Shall You Reap! Toward Preselection Mechanisms for Macrotask Crowdsourcing



Ujwal Gadiraju and Mengdie Zhuang

Abstract Crowdsourcing marketplaces have been flourishing over the last decade, providing a new source of income for hundreds of thousands of people around the globe. Different from microtasks, which are simple and require innate human intelligence in return for small amounts of monetary compensation, the work available on freelancing platforms or in macrotasks often requires a skilled workforce, considerably more time to complete, but the associated rewards are relatively larger and commensurate. Therefore, forming efficient collaboration among workers and finding experts are crucial for ensuring the quality of macrotasks. Worker preselection can be used to ensure that desirable workers participate in available tasks in crowdsourcing marketplaces. In this chapter, we describe two novel preselection mechanisms that have been shown to be effective in microtask crowdsourcing. We discuss how these preselection mechanisms can be used within macrotasks.

6.1 Introduction

The emergence of crowdsourcing of paid work (Howe 2006) has created a global market for online labor where human input can be readily acquired and services can be rendered irrespective of time of the day or location, via a number of platforms (Pongratz 2018). These crowdsourcing platforms vary based on the nature of human input that is required and available, ranging from small human intelligence tasks (HITs) that require a few minutes and no specialized skills (called *microtasks*) to

U. Gadiraju (✉)

L3S Research Center, Leibniz Universität Hannover, Appelstr. 4, 30167 Hannover, Germany
e-mail: gadiraju@L3S.de

M. Zhuang

University College London, London, UK
e-mail: m.zhuang@ucl.ac.uk

© Springer Nature Switzerland AG 2019

V.-J. Khan et al. (eds.), *Macrotask Crowdsourcing*,

Human–Computer Interaction Series, https://doi.org/10.1007/978-3-030-12334-5_6

longer, more complex and often creative tasks that may require specialized skills and several hours to complete (called *macrotasks*).

Typically in a paid microtask crowdsourcing system, a worker accesses the tasks available and chooses which task(s) to complete. The factors that influence a worker's choice in task selection have been studied in detail in previous works (Kaufmann et al. 2011; Gadiraju et al. 2014). The self-centric and subjective nature of task selection on a large crowdsourcing platform (such as Amazon's Mechanical Turk¹ or FigureEight²) is apparent, i.e., it is up to the crowd workers to select a task according to their interests, preference, or expertise. The increasing popularity of crowdsourcing microtasks along with the range of platforms facilitating such efforts, can lead to an overload of choices for a crowd worker. As pointed out by Barry Schwartz in his influential psychology and social theory works, an overload of choices often tends to have detrimental effects on the decision-making process of people (Schwartz and Ward 2004; Schwartz 2004). The large variety of choices in the tasks that are available for an experienced crowd worker (Chilton et al. 2010) makes it difficult for one to select an appropriate task to complete; workers struggle to find tasks that are most suitable for them.

Prominent microtask marketplaces such as AMT or F8 serve as intermediaries to numerous other crowdsourcing channels, by gathering and accumulating large numbers of diverse tasks serving various ends (Kittur et al. 2008; Georgescu et al. 2014). The effort required to search for suitable tasks (in terms of workers' competencies or interests), or in some cases a lack of alternatives (Gadiraju et al. 2014), leads to workers settling for less suitable tasks. The quality of the work thus produced eventually decreases. This is supported by the findings of Chilton et al. (2010), where the authors found that workers most often choose tasks from the first page of the "recently posted tasks", or the first two pages of "tasks with most available instances". More recently, a study of the dynamics of microtasks on AMT by Difallah et al. showed that recently published tasks have almost ten times higher attractiveness for workers as compared to old tasks (Difallah et al. 2015). This skewed attention based on created time or available tasks in one request is independent from workers' experience or expertise. This is supported by the findings of Chilton et al. (2010), where the authors found that workers most often choose tasks from the first page of the "recently posted tasks", or the first two pages of "tasks with most available instances". More recently, a study of the dynamics of microtasks on AMT by Difallah et al. showed that freshly published tasks have almost ten times higher attractiveness for workers as compared to old tasks (Difallah et al. 2015). While some workers settle to work on tasks that are not optimally suited to them, some more capable workers may be deprived of an opportunity to work on the tasks they are ideally suited for, due to limitations on the number of participants or individual contributions. Workers often participate in tasks which are beyond their competence and skills, despite their inherent attempt to maintain their reputation. Thus, the overall effectiveness of the crowdsourcing paradigm decreases. The consequences of such suboptimal *worker-task matching*,

¹AMT—<https://www.mturk.com/mturk/>.

²F8—<http://www.figure-eight.com/>.

include decreased effectiveness of tasks, the possibility of damaging workers' reputation due to their entailing performance, the decreased task availability for more suitable workers, and a level of worker engagement.

Similar phenomena can be observed in case of macrotask marketplaces as well. The process is regulated by participants who can self-select into contests they prefer, and similar worker-task matching issues exist. In crowdsourcing contest platforms like InnoCentive³ and TopCoder,⁴ work is issued as an open call and anyone can participate in any job and the best submission wins the reward (Archak and Sundararajan 2009; DiPalantino and Vojnovic 2009). Even in freelance markets such as Upwork,⁵ in which specialized jobs must be performed by skilled workers (Chatterjee et al. 2015), some macrotasks may consist of multiple steps or require separate skills, and thus collaboration among freelancers are needed. Due to the variety in the skills required to complete available jobs successfully, together with the limited freelancer ability, a given job may have to be divided among freelancers (Ho and Vaughan 2012). To team suitable workers together and create optimal teams is a pivotal task.

It is therefore essential to solve the problem of unsuitable freelancers or workers participating in tasks or teaming-up on the same task on crowdsourcing platforms. To do so, two research areas have been identified to tackle these two problems, namely *expert finding* and facilitating optimal collaboration through *effective team formation*. One can rely on reputation systems in place for selecting freelancers or workers with a desirably high reputation. For example, within freelancing marketplaces, freelancers have a reputation level as well as a minimum acceptable hourly rate, and a set of skills that they possess. Similarly in microtask crowdsourcing platforms, worker reputation can be used as a basis for preselection. However in the context of microtask crowdsourcing, it has been shown that these measures are insufficient to guarantee adequate quality of results (Gadiraju et al. 2015; Kittur et al. 2013). Moreover, several prescreening methods that have been adopted for microtask design, are generally based on the performance of workers on prototypical tasks (Oleson et al. 2011). If a worker passes a prototypical task or a qualification test, then she can proceed to participate in the actual task. This means that the performance of a worker in a prototypical task is assumed to be an indicator of the competence of a worker.

Whether or not such preselection strategies can be effective in macrotask crowdsourcing has remained unexplored. In this chapter, we present two novel preselection mechanisms which have been shown to be effective in microtask crowdsourcing marketplaces and discuss whether these techniques can yield similar results in case of macrotasks.

³<http://www.innocentive.com/>.

⁴<http://www.topcoder.com/>.

⁵<http://www.upwork.com/>.

6.2 Background and Related Literature

In this section, we discuss related literature from different fields; we first elaborate on relevant work in the space of (i) macrotask crowdsourcing, and (ii) self-assessment, to provide the context of this chapter. We then introduce recent works on (iii) competence of crowd workers, and (iv) quality control in crowdsourcing.

6.2.1 *Macrotask Crowdsourcing*

Cheng et al. defined *macrotasks* as large tasks that require relatively more time to complete (for example, transcribing a speech) as opposed to microtasks that are easier to complete and require lesser time (for example, transcribing a single sentence from a speech) (Cheng et al. 2015). The authors compared macrotasks to microtasks and found that (a) decomposing macrotasks into microtasks resulted in longer task completion times, but higher quality outcomes, and that (b) workers indicated a better experience that is more resilient to interruptions in case of microtasks. In work by Haas et al., the authors referred to context-heavy data processing tasks that require many hours of work as *macrotasks* (Haas et al. 2015). They argued that macrotasks represent a trade-off between microtasks and freelance knowledge work; wherein they provide the automation and scale of microtasks, while supporting much of the complexity of traditional knowledge work at the same time. In this context, the authors present “Argonaut”, a framework that extends existing data processing systems by facilitating the use of high-quality crowdsourced macrotasks. The framework presents the output of automated data processing techniques as the input to macrotasks and instructs crowd workers to eliminate errors, leading to significant performance gains. More recently, Schmitz and Lykourantzou proposed a model that supports the sequential improvement of a given macrotask one worker at a time, across distinct time slots of a given timeline, until a sufficient quality level is achieved (Schmitz and Lykourantzou 2018). This lies in contrast to splitting a macrotask into several microtasks and assigning them to workers in parallel.

In this chapter, we reflect on the existing novel methods for worker preselection that have been shown to be effective in crowdsourced microtasks. We discuss attributes of these methods that can render them suitable for worker preselection in macrotask crowdsourcing.

6.2.2 *Self-assessment*

The Dunning-Kruger effect is a cognitive bias that entails inflated self-assessment and illusionary superiority amongst incompetent individuals (Dunning 2011). The authors proposed that incompetence in a particular domain reduces the metacognitive

ability of individuals to realize it. Skills that encompass competence in a particular domain are often the same skills that are necessary to evaluate competence in that domain. For example, consider the ability to solve a Math problem; the skills required to solve the problem are the same skills that are necessary in order to assess whether the Math problem has been accurately solved. The authors attribute this bias to the metacognitive inability of incompetent individuals. On the other hand, competent individuals tend to underestimate their relative competence due to falsely assuming that tasks that they find easy are also easy for others. The authors thereby show that incompetent individuals cognitively miscalibrate by erroneously assessing oneself, while competent individuals miscalibrate by erroneously assessing others.

Apart from the work of Kruger and Dunning (1999), there have been several other noteworthy works in the realm of individual self-assessment. Research works have shown that people provide inflated self-evaluations on performance in a number of different real-world settings. Dunning et al. showed and discussed the implications of such flawed self-assessments on health, educational settings, and the general workplace (Dunning et al. 2004).

Kulkarni et al. showed that in an online course addressing a large number of students (MOOC), the students graded their work 7% higher than those assigned by the staff on average (Kulkarni et al. 2015). Other existing data from experiments reinforce the mistaken self-evaluation of performance (Ehrlinger and Dunning 2003; Ehrlinger et al. 2008). These works show that incompetent individuals are worse at assessing the quality of performance and often tend to think that they outperform the majority, while in fact they belong to the lower rungs of the performance quartile. Complementing these existing works on self-assessment, in our work we aim to understand whether the flawed self-assessment theories hold among crowd workers in the crowdsourcing paradigm. In contrast to these studies that are largely based on self-selected groups of individuals leading to potential selection bias, we use the crowd as a source for a diverse landscape of individuals with respect to their demographics, skills, and competence.

Despite a considerable number of works that assert the findings from the Dunning-Kruger effect, the underlying reasons that dictate the dual-curse resulting in the miscalibrated self-assessment have been widely contested (Burson et al. 2006; Krajc and Ortmann 2008; Krueger and Mueller 2002). Several researchers have provided alternative accounts for the Dunning-Kruger effect, alluding it to regression to the mean and the above-average effect. These accounts have in turn resulted in rigorous theoretical responses and empirical refutations (Ehrlinger et al. 2008), and are out of the scope of our work in this chapter.

In closely related work that proposes the use of self-assessments to improve crowd work, Dow et al. showed that self-assessments allowed workers to improve over time in a task involving writing consumer reviews of products they owned (Dow et al. 2012). The authors of this work proposed the use of self-assessments to yield better work quality by promoting self-reflection and learning. In contrast, we propose to consider the accuracy of worker self-assessments alongside their task accuracy in a preselection phase as an indicator of their true competence and potential per-

formance. Thus, we develop a distinct and novel approach by directly leveraging self-assessments as a worker filtering mechanism, rather than aiming to improve work through self-review.

6.2.3 *Competence of Crowd Workers*

The crux of prior research works in the realm of characterizing crowd workers has mainly focused on ensuring reliability of workers, and presenting a means to the requester to preselect prospective workers (Kittur et al. 2013). In this regard, researchers have suggested the use of prescreening methods and qualification tests (Kazai 2011), trust models to predict the probability of reliable responses (Yu et al. 2012), hidden gold standard questions (Oleson et al. 2011), and the use of metrics that quantify acceptability of responses from the crowd (Gadiraju et al. 2015b). In this chapter, we propose a novel method for the preselection of workers, that outperforms traditional performance-based prescreening methods.

Kazai et al. (2011) used behavioral observations to typecast workers as one of *Spammer*; *Sloppy*; *Incompetent*; *Competent*; or *Diligent*. Here the authors take a keen interest in designing this typology with an aim to attract workers with desirable features, rather than to understand the competencies of the worker population.

As discussed by Dukat and Caton (2013), these existing approaches are seldom applied to ascertain actual worker competencies. They merely serve as an indicator for whether a worker is likely to possess the required ability to complete a microtask successfully, and whether a worker is trustworthy. In this chapter, we present an understanding of the diversity in competence of individual crowd workers.

In closely related works by Kosinski and Bachrach et al., the authors measured the performance of crowd workers on a standard IQ questionnaire (Bachrach et al. 2012; Kosinski et al. 2012). The authors however, discuss factors that effect the overall performance such as composition of the crowd, reputation of workers and monetary rewards. Finally, the authors discuss an approach to aggregate responses from crowd workers to boost performance. While in these works the authors show that aggregating responses from crowd workers is a profitable approach, in this chapter, we are more interested in the individual competence of workers, and therefore adopt a more granular view of responses.

Previous works have highlighted the importance of building tools that support crowd work from the perspective of workers, in order to address the power asymmetry in existing crowdsourcing platforms such as AMT (Gadiraju et al. 2017b; Irani and Silberman 2013; Martin et al. 2014, 2016). In addition to this, Kittur et al. identified *facilitation of learning* as an important next step toward building a bright future for crowd work (Kittur et al. 2013), and others proposed methods to improve learning in crowd work (Gadiraju and Dietze 2017). Complementary to these initiatives, we propose the use of self-assessments in preselection of workers to aid requesters in recruiting the desired crowd. In the future, we can explore the potential use of self-assessments to help workers increase their self-awareness, identify and potentially facilitate learning where their skills are lacking. Thus, we believe that there can be

promising new directions based on leveraging workers' self-assessments to support and improve crowd work in various domains.

6.2.4 Quality Control in Crowdsourcing

One of the classic approaches to detect low quality work, is to compare worker responses against a gold standard dataset. Oleson et al. proposed the programmatic creation of gold standard data to provide targeted training feedback to workers and prevent common scamming scenarios. Authors found that it decreases the amount of manual work required to manage crowdsourced labor while improving the overall quality of the results (Oleson et al. 2011). Similarly, Wang et al. proposed to seamlessly integrate gold data (i.e., data with priorly known answers) for learning the quality of workers (Wang et al. 2011).

Another traditional way to increase label quality generated by means of microtask crowdsourcing is to rely on redundancy; by assigning the same task to a number of workers and then aggregating their responses. Sheshadri and Lease have been benchmarked such techniques over a set of crowd generated labels, comparing state of the art methods over the classic majority vote aggregation method (Sheshadri and Lease 2013). More recently, Venanzi et al. proposed an advanced response aggregation technique that weights crowd responses based on measures of workers similarity, showing a significant improvement in label accuracy (Venanzi et al. 2014).

Other works in microtasks crowdsourcing moved their focus from the outcomes of the tasks to crowd workers' behavior and their background. Marshall and Shipman proposed the use of psychometric tests to ensure reliability of responses from workers (Marshall and Shipman 2013). Rzeszotarski and Kittur looked at worker tracking data with the purpose of distinguishing between high and low performing workers (Rzeszotarski and Kittur 2011). Additionally, the authors presented visual analytics tools that allow requesters to observe worker performance and identify low performers to be filtered out (Rzeszotarski and Kittur 2012). Regrading workers' background, Kazai et al. proposed to look at worker demographics and personality traits as indicators of work quality (Kazai et al. 2012). Qualification tests and pre-screening methods have also been adopted in order to select appropriate workers for a given task. Recent work by Gadiraju et al. has proposed the use of worker self-assessments for preselection (Gadiraju et al. 2017a). This approach has the potential to be applied to the macrotasking scenario since some macrotasks may require more than one person to collaborate on the task, or do not have a single "golden answer".

Similarly, workers' personality and their background were studied for preselection in macrotasks, and mostly for forming the right team. Lykourantzou et al. (2016) used the DISC test (Marston 2013) which covers four distinct personality types: *dominance, inducement, submission, compliance*. They observed that in a creative advertisement design task that requires five people, the groups with more balanced personality achieved a significantly better performance, better communication, and higher satisfaction than the imbalanced ones (Lykourantzou et al. 2016). A team

dating mechanism was also introduced in other work (Lykourantzou et al. 2017), where crowd workers interact on brief tasks and rate each other before choosing a partner for longer and more complicated tasks. Such a mechanism allows the worker to select their partner based on how they evaluate their dates, or the person's average rating and results in improvements while performing a creative task (Lykourantzou et al. 2017).

Quality control in crowdsourced macrotasks can also be tackled by eliminating the risk of interruptions, such as breaking macrotasks down into several microtasks. This is because a single macrotask requires longer completion time than a microtask and is thus more vulnerable. We have discussed earlier that task decomposition of a macrotask into several microtasks results in an overall increase in the task completion time and a better performance, and is especially suitable for environment with interruptions (Cheng et al. 2015). In this chapter, we focus on worker self-assessment and behavior rather than personality or background using self-assessment.

A limitation of prior works on quality control based on worker typologies is the absence of prior knowledge about worker types in typical scenarios, and the lack of automated methods that go beyond identifying *good* and *bad* performing workers. Our work is complementary to aforementioned prior works, in that we aim to improve the quality of work that is produced by workers. In addition, by relying on a more granular understanding of worker types, we afford preselection of desired workers in the absence of any prior information about workers. We extract behavioral features and propose a supervised machine learning model, that automatically detects worker types, thus going beyond the good/bad binary classification problem.

In the next section, we present a worker self-assessment based prescreening mechanism that has been validated in two major types of microtasks, and discuss the potential application of this method for preselection in macrotask crowdsourcing. Following that, we describe a behavior-based worker preselection mechanism with the notable advantage of good performance in the absence of gold questions, and the potential usage in macrotasks.

6.3 Operationalizing Worker Self-assessments

Through rigorous experiments we found evidence of the existence of the Dunning-Kruger effect in the crowd (Gadiraju et al. 2017a). We found that not all crowd workers are adept at making accurate self-assessments; competent workers are relatively better at doing so. This is further reinforced by our findings in a tagging task, where we observed that competent workers produced tags with both higher quality as well and quantity (cf. Gadiraju et al. 2017a). Based on this understanding, we propose that it can be beneficial to operationalize worker self-assessments as an indicator of worker competence and therefore performance. We choose to use accuracy of worker self-assessments in the prescreening tasks in addition to their actual performance in the prescreening tasks to select workers. Thus, the only additional requirement in our proposed method is a self-assessment question at the end of the pre-screening tasks,

making it straightforward to implement. Figure 6.4 illustrates the traditional prescreening method (Fig. 6.4a) in comparison to our proposed self-assessment based prescreening approach (Fig. 6.4b).

6.3.1 Evaluation in Sentiment Analysis Task

From our earlier findings (cf. Gadiraju et al. 2017a) we note that some crowd workers (less-competent) exhibit inflated self-assessments. We also found that the competent workers produce significantly better quality of work, as observed in a tagging task. Next, we seek to answer whether we can operationalize the ability of workers to accurately self-assess their performance in a real-world microtask, in order to preselect a more suitable crowd with respect to the task. Can worker self-assessments be used as a means to provide a stronger indicator of worker competence?

We evaluated our proposed method of using worker self-assessments as a basis for prescreening crowd workers, as opposed to traditional prescreening that is purely based on the performance of workers. We considered a popular crowdsourcing task; *sentiment analysis* (Gadiraju et al. 2014). In this task composed of 30 units, crowd workers are asked to read a tweet in each unit and classify the projected sentiment as either `positive`, `negative` or `neutral`. For this purpose we use the dataset introduced by Gadiraju et al. (2015a), that consists of expert-classified tweets, thereby providing our ground truth. Although the monetary compensation for two prescreening methods were customized differently, we compensated the participating workers with an hourly wage of over 7.5 USD in both cases.

Self-Assessment Based Prescreening We prototyped a 5-unit task for the sentiment analysis, consisting of tweets different from those in the actual 30 units considered for the evaluation task. On completing these 5 units, workers are asked the question, “*How many questions do you think you answered correctly?*”. We consider a worker to have passed this screening task, if the worker accurately predicts her score while the actual score is more than 3 out of 5, or if the worker miscalibrates her prediction by one point while her actual score is more than 3 out of 5 (i.e., *miscalibration* = 0 or 1). The intuition behind using a threshold of “3” is due to our aim to replicate a realistic preselection scenario. FigureEight suggests a minimum accuracy of 70% by default⁶ for the traditional prescreening method (which is actual score more than 3 out of 5 in our case).

We deployed this task on FigureEight and gathered responses from 300 workers by offering a compensation of 2 USD cents. We found that only 110 out of 300 workers passed the threshold of actual score more than 3 out of 5. Of these 70 workers passed the self-assessment accuracy criteria and thereby passed the prescreening. Next, we deployed the actual evaluation task consisting of 30 units to these 70 workers alone⁷

⁶As per FigureEight’s guide to test questions and quality control.

⁷FigureEight provides support for this via the *internal workforce*.

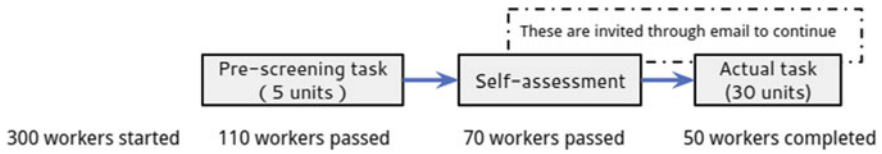


Fig. 6.1 Flowchart depicting the recruitment of workers and their progress through the *self-assessment based prescreening* evaluation

by using their e-mail IDs. We offered a reward of 5 USD cents to workers. Within a span of 1 week, 50 of the 70 workers completed the task as shown in Fig. 6.1.

Traditional Prescreening One week later, we deployed an identical task consisting of the same 30 units on FigureEight. There was no overlap in the pool of workers across the two tasks. Hence, the observed results are not due to ordering effects. We used the same 5 units in the traditional prescreening process as in the case presented above, and only those workers who answered more than 3 out of 5 units correctly were allowed to participate in the actual task. We gathered responses from 50 distinct workers, and these workers were also paid a compensation of 5 USD cents (to match the incentive offered and number of collected judgments in the self-assessment based prescreening method).

Results We evaluated the two different methods based on the following two aspects: accuracy of the preselected workers in the tasks following the screening, and their task completion time. We found that the self-assessment based prescreening method (green dots in Fig. 6.2) resulted in workers who performed with an accuracy of nearly 94% on average, with an inter-annotator agreement of 0.95 (computed by pairwise percent agreement (PPA)). The traditional prescreening method (presented in Fig. 6.2 in the red color) resulted in workers who performed with an average accuracy of around 78%, with an inter-annotator agreement of 0.83 (computed by PPA).

We found that the difference in the resulting worker performances between using the self-assessment based prescreening method ($M = 27.95$, $SD = 1.79$) and the traditional prescreening method ($M = 23.63$, $SD = 6.23$) was statistically significant $t(95) = 3.40$, $p < 0.01$, with a large effect size; *Cohen's d* = 0.94. We did not find a significant difference in the task completion time of workers resulting from the two different methods of prescreening.

It is important to note that in the self-assessment based prescreening method, the average actual scores of workers on the qualification test was 4.4/5 and that of workers in the traditional prescreening method was 4.3/5, without a significant difference. This shows that the observed improvement is due to the consideration of worker self-assessments, and not simply a result of selecting workers who performed better in the prescreening phase. We highlight that there may be a confound in having workers wait, then self-select to return and complete the actual evaluation task in the self-assessment based prescreening method. Such workers may be more diligent than workers in the traditional prescreening method, who immediately began the actual evaluation task. However, due to the number of workers in the pool, the significant

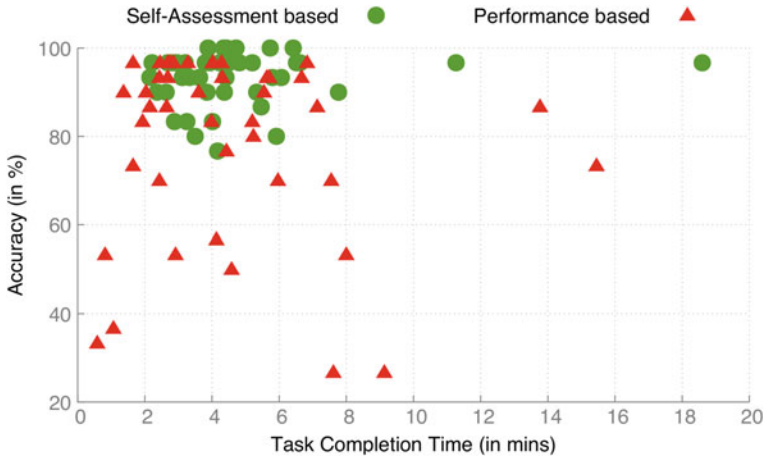


Fig. 6.2 Performance of workers acquired by the proposed *self-assessment based prescreening* and by traditional *performance based prescreening*

differences and the large effect size observed, we believe this does not risk the overall result and does not pose a threat to its validity.

From these results, we observe that prescreening crowd workers based on their self-assessments provide a better reflection of their actual competence, leading to an improved quality of results. We note an improvement of over 15% in accuracy and 12% in agreement between workers by using self-assessment based prescreening of workers in a sentiment analysis task. Thus, we can conclude that operationalizing self-assessments of workers in a given task in conjunction to their performance in the task, can serve as a stronger indicator of worker competence than relying on worker performance alone.

6.3.2 Evaluation in Verification and Validation Task

We operationalized worker self-assessments in a sentiment analysis task earlier and improved the preselection of crowd workers. Similar to the sentiment analysis task described in the previous section, we considered an additional real-word task of image validation. Our aim is to verify whether our proposed approach would yield similarly improved results in another type of task, due to the effectiveness of our proposed worker preselection method.

In this task composed of 13 units in total, crowd workers were asked to analyze the pictures in online automobile ads to spot mismatched information. To publish an online ad, sellers need to textually describe the state of the vehicle (damaged or not) and its mileage. Sellers commonly omit damage-related information from the description or claim a lower mileage in order to achieve a better placement in the



(a) Seller declared visible damage in the description of the advertisement.

(b) Seller omitted visible damage-related details from the description of the advertisement.

Fig. 6.3 Example automobile ads from the online marketplace *mobile.de* that either **a** declare damages in the vehicle description, or **b** omit damage-related information

search results (see Fig. 6.3). In many cases this information is evident in the pictures. While this cannot be easily detected by automated algorithms, it is a rather simple task for humans.

Task Design We used a within-subjects design, and manually found and annotated a total of 13 vehicle ads⁸ which served as groundtruth for the task. Each ad corresponds to one unit where workers are asked to answer three multiple choice questions: (i) Is the car marked as damaged? (ii) Can you identify that the car has a visible damage or functional problems based on the pictures? (iii) Is the mileage information consistent with the picture? We took care to find distinct ads that produced an even distribution of the options corresponding to each question. The units were randomized and after answering 3 units (total of 9 questions), workers were asked to assess their performance on the 9 questions. With an aim to compare self-assessment based prescreening with performance-based prescreening, all workers were allowed to continue onto 10 more units. Each worker was rewarded with 5 USD cents on successful task completion, which is more than 7.5 USD per hour. We deployed this task on FigureEight and collected responses from 100 distinct workers.

Results: Traditional Prescreening Similar to the previous sentiment analysis task, the traditional prescreening method is characterized by a performance threshold of 70% in the prescreening phase. Thus, we filtered out workers (36 in total) who did not achieve a minimum of 70% accuracy in the first 3 units (9 questions). In the 10 units that followed, comprising the actual task, this group of workers ($N = 64$) achieved an average accuracy of 84.05% ($M = 84.05$, $SD = 10.35$), with an inter-annotator agreement of 0.81 using pairwise percent agreement (PPA).

Results: Self-Assessment Based Prescreening In case of the proposed self-assessment based prescreening approach, we consider the accuracy of worker self-

⁸We used publicly available ads from the online marketplace <http://www.mobile.de/>.

assessments in addition to the 70% accuracy threshold in the prescreening phase. Here again, we tolerate an error of 1 point in the workers self-assessments (i.e., *miscalibration* = 0 or 1). Workers who passed this prescreening phase ($N = 49$), performed with an accuracy of 89.6% ($M = 89.6$, $SD = 6.6$) in 10 units that followed, comprising the actual task. In this case, the inter-annotator agreement was found to be 0.9 (PPA).

To summarize, we found that 64 of the 100 workers passed 70% accuracy threshold. Of these, 49 workers passed the self-assessment accuracy criteria and thereby passed the prescreening. The self-assessment based prescreening approach resulted in an improvement in accuracy of nearly 6%, and an increase in the inter-annotator agreement between workers by 8% in comparison to the traditional prescreening method. The difference in worker accuracy between the traditional and the self-assessment based prescreening methods was found to be statistically significant with a moderately large effect size; $t(112) = 2.60$, $p < 0.01$, Hedge's $g = 0.62$. Once again, we noted that the difference in performance in the prescreening phase (3 units, 9 questions) across the two groups of workers was not statistically significant, indicating that the improvement in the accuracy of workers using our proposed approach is due to the consideration of accuracy of workers' self-assessments. We also did not find a significant difference in the task completion time of workers selected using the different methods.

6.3.3 Why Self-assessment Based Preselection Can be Beneficial in Macrotask Crowdsourcing?

In the preceding section, we described the operationalization of worker self-assessments as a preselection mechanism and the significant benefits of employing this method in two different types of microtasks (sentiment analysis task, verification and validation task). Here, we aim to address the question of whether we can apply this mechanism for macrotask crowdsourcing. It must be noted that operationalizing self-assessments may not always be straightforward when it comes to macrotasks. The difficulty in applying self-assessment based preselection to macrotasks, stems from the need for gold questions, which not only have to be relatively short to complete but also cover the same skills required by the main tasks. Due to the nature of macrotasks, it may be challenging to create such gold questions (e.g., creative tasks) or it may be difficult to create a small portion of questions that may cover all the desirable skills (e.g., the pursuit of a contest). However, if these hurdles can be overcome, self-assessment based prescreening can be equally effective for macrotask crowdsourcing since the underlying meta-cognitive principles that govern the mechanism would remain the same.

The self-assessment based preselection mechanism is worthwhile for the macrotasks that can be divided into smaller ones that require the same skills. One popular example is translation. The feasibility of dividing such tasks have been demonstrated

in Cheng et al. (2015), and using preselection in such a task has the potential to improve the quality of the results.

Even for the tasks that require multiple skills and collaboration, we could identify the individual roles with required skills, which is a subset of all the skills required, and design a set of gold questions to represent each role. In other words, reducing the number of skills that need to be covered in a batch of gold questions. It is much like producing job descriptions and testing questions for employees that have different skills to work on the same project. Consider the example of creating a broad set of analytic reports with multimedia (a popular task type on Fiverr⁹). This specific type of macrotasks require analyzing data, designing the main thread, article writing, and animation creation. There are a larger number of freelancers or crowd workers that have a subset of the skills than the ones that fulfill all the skills. Therefore, it should be feasible to use self-assessment based preselection for each individual skill to recruit a team that can complete the required work effectively.

6.4 Behavior-Based Worker Preselection

In the previous section, we described how accuracy of worker self-assessments can be used together with their accuracy in a prescreening test for effective preselection of workers. Such a method requires the worker to respond to a single self-assessment question. Next we present an unobtrusive method based on data that is collected during task completion in a short prescreening phase. In this section, we describe a method to preselect workers based on their behavioral type (defined according to a worker typology). Figure 6.4 illustrates the traditional prescreening method (Fig. 6.4a) in comparison to our proposed worker behavior-based prescreening method (Fig. 6.4b).

6.4.1 Why a Worker Behavior-Based Method?

Rzeszotarski and Kittur, proposed to track worker activity to distinguish between *good* and *bad* workers according to their performance (Rzeszotarski and Kittur 2011). Recently, Dang et al. built a framework called *mmm*Turkey, by leveraging this concept of tracking worker activity (Dang et al. 2016). Rzeszotarski et al. showed several benefits of their approach when compared to other quality control mechanisms due to aspects such as effort, skill, and behavior that can be interpreted through a worker's activity, and eventually help in predicting the quality of work (Rzeszotarski and Kittur 2011, 2012). While it is certainly useful to predict good versus bad quality of work, we argue that further benefits can be revealed by understanding worker activity at a finer level of granularity. For example, the knowledge that even *good* workers

⁹<https://www.fiverr.com/>.

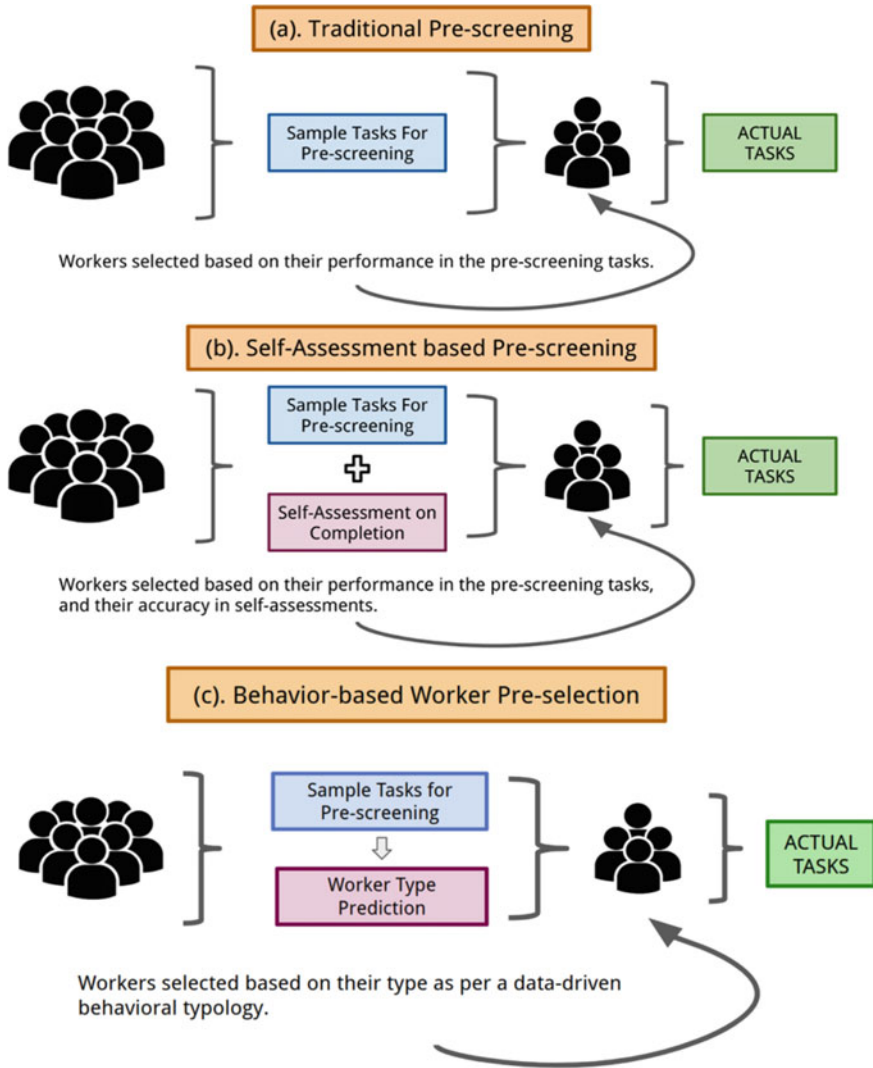


Fig. 6.4 Comparison between **a** the traditional prescreening method based on worker performance in prescreening tasks, **b** the self-assessment based prescreening method which considers worker performance in the prescreening tasks as well as their accuracy in self-assessments, and **c** the worker type based prescreening method which considers the behavioral traces of workers to select desired types of workers

perform and operate in different ways to accomplish tasks, leads to the question of whether such differences can have practical implications.

6.4.2 Modeling Worker Behavior

We first present a worker typology by building on prior works in an inductive and data-driven fashion prescribed by Berg et al. (2004). We collected data from 1800 workers (1800 HITs in total) completing Information Finding (IF) and Content Creation (CC) tasks each. Overall, we compensated workers at an hourly wage of 7.5 USD. In the latter half of this section, we introduce the low-level behavior features indicating workers' behavioral traces, as we expect these to be informative in predicting worker types.

Worker Typology To summarize, Kazai et al. (2011), Gadiraju et al. (2015b), and Vuurens and De Vries (2012) proposed worker typologies based on worker behavior and performance, while Eickhoff et al. (2012) categorized workers based on their motivation. We propose to combine behavior, motivation, and performance based on the works mentioned above (Kazai et al. 2011; Gadiraju et al. 2015b; Vuurens and De Vries 2012), rather than looking at each aspect individually to typecast workers from a holistic standpoint. Based on the responses provided by workers in 1,800 HITs, we computed their performance. As described by Eickhoff et al. (2012), money-driven workers are motivated by the monetary incentives, while entertainment-driven workers mainly seek diversion but readily accept the monetary rewards as additional extrinsic motivation. Thus, we explicitly asked workers about their motivation for participation at the start of task. Finally, based on the low-level worker activity that we logged, we were able to analyze worker behavior.

To categorize workers based on their performance (accuracy and task completion time), motivation, and behavior we used a data-driven and inductive approach. This means that the categories we thereby derived were grounded in the data from which they emerged, as suggested by Denzin (1978) as well as Glaser and Strauss (1967). We manually inspected workers' responses to the 1,800 HITs and built rubrics around their task completion time, trustworthiness, and performance to assign appropriate labels. The rubrics were such that worker types could be assigned without clashes between the classes. We designed a coding frame according to which we could decide which category in the typology a worker belonged to. In case, the characteristics exhibited by workers did not fit any existing category, a new one was created. After resolving disagreements on the coding frame every worker was labeled with a category. We followed the guidelines suggested by Strauss (1987), Berg (2004) while conducting the open-coding of behavioral data, collected over the 1,800 HITs run on FigureEight, leading to the following categories.¹⁰ We also describe the rubrics used to categorize workers into the respective category. Table 6.1 summarizes the worker types.

– **Diligent Workers (DW)**. These crowd workers may be *money-driven* or *entertainment-driven*. They make sure to provide high-quality responses and spend a long time to ensure good responses.

¹⁰Note that worker types describe session-level behavior of the workers rather than properties of a person.

Rubric used to categorize DW: trustworthy workers who have high to very high task completion times (i.e., 3rd and 4th quartiles of task completion times among all workers in the given task), and high to very high accuracy (i.e., 3rd and 4th quartiles of accuracy among all workers in the given task).

– **Competent Workers (CW).** These crowd workers may be *money-driven* or *entertainment-driven*. They possess skills necessary to complete tasks in a quick and effective manner, producing high quality responses.

Rubric used to categorize CW: trustworthy workers who have very low to low task completion times (i.e., first 2 quartiles of task completion times among all workers in the given task), and high to very high accuracy (i.e., 3rd and 4th quartiles of accuracy among all workers in the given task).

– **Fast Deceivers (FD).** These crowd workers are *money-driven*, and attempt to complete a given task in the fastest possible way to attain the rewards offered. Due to this, *fast deceivers* provide poor responses by copy-pasting content and taking advantage of loopholes in the task design (such as weak or missing validators).

Rubric used to categorize FD: untrustworthy workers¹¹ who have low to very low task completion times (i.e., first 2 quartiles of task completion times among all workers in the given task), and very low accuracy (i.e., the bottom quartile of accuracy among all workers in the given task).

– **Smart Deceivers (SD).** These crowd workers are *money-driven* and aware of potential validators and checks that task requesters may be using to flag workers (such as minimum time spent on a question). They provide poor responses without violating validators, and thereby exert less effort to attain the incentives.

Rubric used to categorize SD: trustworthy workers who have high task completion times (i.e., 3rd quartile of task completion times among all workers in the given task), and very low accuracy (i.e., the bottom quartile of accuracy among all workers in the given task).



– **Rule Breakers (RB).** These crowd workers may be *money-driven* or *entertainment-driven*. They provide mediocre responses that fall short of the expectations of a requester (e.g., providing 3 keywords where 5 are required).















Rubric used to categorize RB: trustworthy workers who have high task completion times (i.e., 3rd quartile of task completion times among all workers in the given task), and high accuracy (i.e., the 3rd quartile of accuracy among all workers in the given task).

– **Less-competent Workers (LW).** These crowd workers may be *money-driven* or *entertainment-driven*. They appear to have a genuine intent to complete a given task successfully by spending ample time on it, but lack the necessary skills to provide high-quality responses.

Rubric used to categorize LW: trustworthy workers who have very high task completion times (i.e., 4th quartile of task completion times among all workers in

¹¹Untrustworthy workers are those workers who failed to pass at least one attention check question.

Table 6.1 Worker types and their associated motivation, and rubrics (task completion time, accuracy). The  represents the ordered four quartiles, in which the 1st quartile on the left and the 4th quartile on the right. The black block  represents the rubrics used to categorize the work type

Type	Motivation	Completion Time	Accuracy
DW	Money/Entertainment	 ;	 ;
CW	Money/Entertainment	 ;	 ;
FD	Money	 ;	 ;
SD	Money	 ;	 ;
RB	Money/Entertainment	 ;	 ;
LW	Money/Entertainment	 ;	 ;
SW	Money/Entertainment	 ;	 ;

the given task), and low accuracy (i.e., the 2nd quartile of accuracy among all workers in the given task).

– **Sloppy Workers (SW).** These crowd workers may be *money-driven* or *entertainment-driven*. They complete tasks quickly and perform with an average or below average accuracy. Sloppy workers (Kazai et al. 2011) appear to err due to their speed within the task.

Rubric used to categorize SW: trustworthy workers who have very low task completion times (i.e., first quartile of task completion times among all workers in the given task), and low accuracy (i.e., the 2nd quartile of accuracy among all workers in the given task).

Features Indicating Behavioral Traces We studied the mousetracking data (including keypresses) generated by crowd workers in 1,800 HITs through 9 content creation (CC) and 9 information finding tasks (IF) (cf. Gadiraju et al. 2018), in order to determine features that can help in the prediction of a worker type. We implemented mousetracking using Javascript and the JQuery library, and logged user activity data ranging from mouse movements to keypresses. We took measures to distinguish between workers that use a mouse and those who use a touchpad. We also distinguish between worker mannerisms with respect to scrolling behavior; use of scrollbar as opposed to the mousewheel. In this way, we gathered worker activity data from each of the experimental tasks deployed on CrowdFlower. Apart from this data, we use a Javascript implementation of browser fingerprinting (Eckersley 2010) in order to identify workers that participate in tasks multiple times (“repeaters”) by virtue of using different worker-ids (Gadiraju and Kawase 2017). We take measures to avoid privacy intrusion of workers by hashing various browser characteristics such as the user agent, cookies settings, screen resolution, and so forth, results in a 64-bit browser fingerprint. We do not retain any worker-specific browser traits other than the resulting fingerprint to identify repeaters. Some of the important features are presented below. A complete list of features used can be found here.¹²

¹²Shortened URL—<https://goo.gl/jjv0gp>.

- time: The task completion time of a worker.
- tBeforeLClick: The time taken by a crowd worker before responding to the multiple choice demographic questions in the tasks.
- tBeforeInput: The time taken by a crowd worker before entering a transcription in the content creation task or a middle-name in the information finding task.
- tabSwitchFreq: Number of times that a worker switches the tab while working on a particular task.
- windowToggleFreq: Number of times that a worker toggles between the current and last-viewed window while working on a particular task.
- openNewTabFreq: Number of times that a worker opens a new tab while working on a particular task.
- closeCurrentTabFreq: Number of times that a worker closes the current tab while working on a task.
- windowFocusBlurFreq: Number of times that the window related to the task goes in and out of focus until task completion by the crowd worker.
- scrollUp/DownFreq: Number of times that a worker scrolls up or down while working in a task respectively.
- transitionBetweenUnits: Number of times a worker moves the cursor from one unit to another in the task.
- totalMouseMoves: The total number of times that a worker moves the cursor within the task.

6.4.3 *Evaluation in Information Finding and Content Creation Tasks*

By exploiting the expert annotated HITs and the features defined based on worker behavioral traces described earlier, we first train and test a random forest classifier to predict worker types at the end of a completed task. Then, we further evaluate the worker type predictions by using it as a preselection criteria and comparing it against the standard qualification test. We study the effectiveness of our supervised models using the same dataset described in the previous section and to predict worker type in CC and IF tasks with varying task complexity. We had 100 different workers for each of the 9 difficulty levels and task lengths combination (3 difficulty levels, 3 different task lengths). We distinguish models by training with two different sets of behavioral features that does or does not contain “gold questions” information of task (i.e., questions with known answers used to check for work quality, which is necessary in the standard qualification test).

Predicting Worker Types Tables 6.2 and 6.3 present Accuracy and F-Measure (to account for unbalanced classes) of our supervised worker type classifiers evaluated using 10-fold cross validation over IF and CC tasks.

We can observe that it is easier to predict worker types when gold questions are available in the task. We also observe higher accuracy of automatic worker type

Table 6.2 Supervised worker type classification evaluation for IF tasks with varying task complexity

HIT length	With gold questions		W/out gold questions	
	Accuracy	F-Measure	Accuracy	F-Measure
10	77.3	0.748	73.6	0.679
20	74	0.701	74	0.691
30	81.4	0.786	79.8	0.763
HIT difficulty	Accuracy	F-Measure	Accuracy	F-Measure
Level-I	82.3	0.779	80.5	0.754
Level-II	79.4	0.77	74.6	0.718
Level-III	72.3	0.691	64.2	0.587

Table 6.3 Supervised worker type classification evaluation for CC tasks with varying task complexity

HIT length	With gold questions		W/out gold questions	
	Accuracy	F-Measure	Accuracy	F-Measure
20	69.02	0.671	58.6	0.532
30	84.5	0.828	75.6	0.712
40	80.3	0.768	78.7	0.729
HIT difficulty	Accuracy	F-Measure	Accuracy	F-Measure
Level-I	74.7	0.714	70	0.643
Level-II	77.5	0.746	67.4	0.611
Level-III	72.5	0.696	64.5	0.59

classification for IF in comparison to CC tasks. Moreover, as *longer* tasks typically provide more behavioral signals, they lead to better automatic classification of workers in our typology. A similar conclusion can be drawn for *less difficult* tasks where worker types can be better distinguished. Due to the imbalance in the different worker types, we also ran undersampling and oversampling experiments, that yielded similar results.

Additional results from the supervised classification evaluation showed that the easiest worker types to be predicted are CW (91% accuracy) and DW (87% accuracy) for CC tasks and DW (88.7% accuracy) and FD (86.6% accuracy) for IF tasks. Most confused worker types by our models are SW classified as CW for CC tasks and CW classified as DW for IF tasks. Feature selection by Information Gain shows that the most predictive features to automatically predict the worker type are mouse movement, windows focus frequency, the task completion time, the score, and tipping point¹³ computed from gold questions (when available).

¹³First point at which a worker provides an incorrect response after having provided at least one correct response (Gadiraju et al. 2015b).

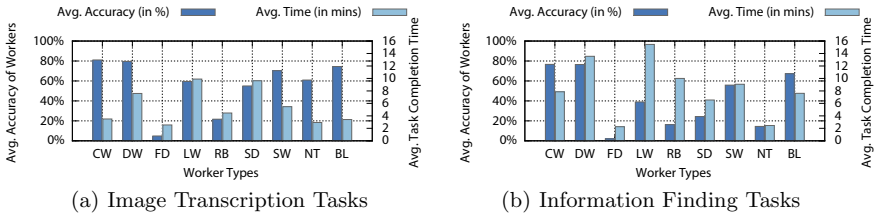


Fig. 6.5 Average accuracy (scaled on the y-axis) and task completion time (scaled on the y2-axis) of the first 5 judgments received from different automatically predicted worker types in the **a** image transcription and **b** information finding tasks. The different worker types presented here are as follows. CW: Competent Workers, DW: Diligent Workers, FD: Fast Deceivers, LW: Less-competent Workers, RB: Rule Breakers, SD: Smart Deceivers, SW: Sloppy Workers, NT (No Type): First 5 judgments without considering worker type, BL (Baseline): First 5 judgments from workers who passed the standard preselection test

Preselection Based on Worker Types Here, we assess the impact of worker type predictions made by the proposed ML models described earlier. Once again we consider the first 5 judgments submitted by workers of each type (worker type as predicted by the classifier). We compare our proposed worker type based preselection method with the standard approach of using qualification tests which we refer to as the Baseline. In the Baseline method, we consider the first 5 responses from each worker to be a part of the qualification test. Only workers who achieve an accuracy of $\geq 3/5$ in the qualification test are considered to have passed the test. This follows our aim to replicate a realistic prescreening scenario.¹⁴ To compare the Baseline method with our proposed approach of worker type based preselection, we consider the first 5 judgments submitted by workers who passed the qualification test.

Figure 6.5 presents the results of our evaluation for the two task types. In case of the image transcription tasks (Fig. 6.5a) we note that on average across all tasks, CW ($M = 81.03, SD = 8.52$) significantly outperform workers in the No Type setting ($M = 60.9, SD = 18.69$) with $t(8) = 5.04, p < 0.0005$. Interestingly, the task completion time (in minutes) of CW ($M = 3.5, SD = 0.85$) is slightly more than that of No Type ($M = 2.93, SD = 0.48$) with $t(8) = 1.86, p < 0.05$. CW also perform significantly better than the Baseline method ($M = 74.41, SD = 14.06$) with $t(8) = 1.86, p < 0.05$. The differences in task completion time between CW and the Baseline method were not statistically significant, indicating that worker type based preselection of CW can outperform existing preselection methods in terms of quality without a negative impact on the task completion time.

For the information finding tasks (Fig. 6.5b), we note that on average across all tasks CW ($M = 76.59, SD = 11.34$) significantly outperform workers in the No Type setting ($M = 14.44, SD = 23.6$) with $t(8) = 5.04, p < 0.0005$. In addition, we also observe that CW significantly outperform workers that are preselected using the Baseline method ($M = 67.26, SD = 14.92$) with $t(8) = 1.86, p < 0.05$. The

¹⁴FigureEight suggests a min. accuracy of 70% by default.

task completion time (in minutes) of CW ($M = 7.87$, $SD = 3.56$) is not significantly different from that of the Baseline method ($M = 7.62$, $SD = 3.45$).

6.4.4 Why Behavior-Based Worker Preselection can be Beneficial in Crowdsourcing Macrotasks?

We present a worker typology, and the associated preselection mechanism based on worker types inferred by their behavior in a prescreening phase. Such worker type based preselection was evaluated in two microtask types (content creation, and information finding tasks). Note that this method only relies on the behavioral data collected in a very small sample of prescreening tasks (5 microtasks). Thus, it is conceivable to use such a method in a short prescreening phase for worker preselection in macrotasks as well. Although having gold questions leads to a better prediction performance of the model, prior work (Gadiraju et al. 2018) has shown that it is not necessary. Thus, such behavior-based worker preselection is applicable to a wider range of macrotasks compared to the self-assessment preselection mechanism. However, one possible concern is about the size of the training data used for prediction. As we discussed, the rubrics used to label worker types are based on the descriptive statistics of the training sample, which is an approximation of the real distribution of workers' performance on this task. The larger the number of workers for training, the better representation we can have and more precise the prediction of worker types will be. Acquiring a reasonable size of data for an unpopular type of macrotask might be troublesome.

Another direction that future work can pursue is to predict workers' background or working context using their behavior. One example is to investigate the role of workers' behavior in inferring workers' personality, and therefore provide extra information for worker team formation. The improvements of considering workers' personality in collaborative macrotasks have been reported in Lykourantzou et al. (2016). However, the assessment of personality requires the workers' response to a set of question, which might be obtrusive to task completion. Using behavior data to infer personality will alleviate this concern.

In addition, workers' behavior can also be garnered to widen the understanding of workers' working environment, such as interruptions, preferred task length, among other attributes.

6.5 Discussion and Conclusions

We found that worker self-assessments can be effectively operationalized within microtasks to serve as useful indicators of true competence. Evaluation results across two different task types showed a robust improvement in the quality of preselection

when accuracy of worker self-assessments in a prescreening phase was considered alongside worker accuracy. We believe that this method would also be effective in macrotask crowdsourcing with the constraint of requiring a prescreening phase. In comparison to microtasks where prescreening phases can be quite short, it is unclear how long such prescreening phases can be for macrotasks. For example, a worker may be required to complete a task that requires 30 min as a prescreening or qualification test before being allowed to complete the actual macrotask. This would have implications on the associated costs, but can still be a useful trade-off. Another limitation is that it may be relatively more difficult for task requesters to use self-assessments as an instrument in more subjective or creative tasks. Thus, although self-assessments based worker preselection provides interesting opportunities, further experiments are required to ascertain the applicability of this mechanism in the landscape of macrotask crowdsourcing.

On the other hand, while behavior-based preselection mechanisms are tolerant to more subjective prescreening phases, the need for a longer prescreening phase in comparison to microtasks persists. Our findings corresponding to the effectiveness of preselecting workers based on their behavioral types, suggest that such methods can be effective even in case of macrotasks. In fact, since behavioral data is arguably richer in case of longer tasks, it may be possible to develop highly accurate predictive models for worker preselection in macrotask crowdsourcing. In the imminent future, we plan to investigate the application of these preselection mechanisms in different types of macrotasks.

References

- Archak, N., & Sundararajan, A. (2009). Optimal design of crowdsourcing contests. In *ICIS 2009 Proceedings* (p. 200).
- Bachrach, Y., Graepel, T., Kasneci, G., Kosinski, M., & Van Gael, J. (2012). Crowd IQ: Aggregating opinions to boost performance. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems* (Vol. 1, pp. 535–542). International Foundation for Autonomous Agents and Multiagent Systems.
- Berg, B. L. (2004). Methods for the social sciences. *Qualitative research methods for the social sciences*. Boston: Pearson Education.
- Burson, K. A., Larrick, R. P., & Klayman, J. (2006). Skilled or unskilled, but still unaware of it: How perceptions of difficulty drive miscalibration in relative comparisons. *Journal of Personality and Social Psychology*, *90*(1), 60.
- Chatterjee, A., Varshney, L.R., & Vishwanath, S. (2015). Work capacity of freelance markets: Fundamental limits and decentralized schemes. In *2015 IEEE Conference on Computer Communications (INFOCOM)* (pp. 1769–1777). IEEE.
- Cheng, J., Teevan, J., Iqbal, S. T., & Bernstein, M. S. (2015). Break it down: A comparison of macro-and microtasks. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (pp. 4061–4064). ACM.
- Chilton, L. B., Horton, J. J., Miller, R. C., & Azenkot, S. (2010). Task search in a human computation market. In *Proceedings of the ACM SIGKDD Workshop on Human Computation* (pp. 1–9). ACM.

- Dang, B., Hutson, M., & Lease, M. (2016, October 30–November 3). MmmTurkey: A crowdsourcing framework for deploying tasks and recording worker behavior on amazon mechanical turk. In *HCOMP'16. Proceedings of the 4th AAAI Conference on Human Computation and Crowdsourcing (HCOMP): Works-in-Progress Track*, Austin, Texas, USA (pp. 1–3). AAAI Press.
- Denzin, N. K. (1978). *The research act: A theoretical orientation to sociological methods* (Vol. 2). New York: McGraw-Hill.
- Difallah, D. E., Catasta, M., Demartini, G., Ipeirotis, P. G., & Cudré-Mauroux, P. (2015). The dynamics of micro-task crowdsourcing—The case of Amazon MTurk. In *24th International Conference on World Wide Web (WWW)* (pp. 238–247). ACM.
- DiPalantino, D., & Vojnovic, M. (2009). Crowdsourcing and all-pay auctions. In *Proceedings of the 10th ACM Conference on Electronic Commerce* (pp. 119–128). ACM.
- Dow, S., Kulkarni, A., Klemmer, S., & Hartmann, B. (2012). Shepherding the crowd yields better work. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work* (pp. 1013–1022). ACM.
- Dukat, C., & Caton, S. (2013). Towards the competence of crowdsourcees: Literature-based considerations on the problem of assessing crowdsourcees' qualities. In *2013 Third International Conference on Cloud and Green Computing (CGC)* (pp. 536–540). IEEE.
- Dunning, D. (2011). The dunning-kruger effect: On being ignorant of one's own ignorance. *Advances in Experimental Social Psychology*, 44, 247.
- Dunning, D., Heath, C., & Suls, J. M. (2004). Flawed self-assessment implications for health, education, and the workplace. *Psychological Science in the Public Interest*, 5(3), 69–106.
- Eckersley, P. (2010). How unique is your web browser? In *Privacy Enhancing Technologies* (pp. 1–18). Springer.
- Ehrlinger, J., & Dunning, D. (2003). How chronic self-views influence (and potentially mislead) estimates of performance. *Journal of Personality and Social Psychology*, 84(1), 5.
- Ehrlinger, J., Johnson, K., Banner, M., Dunning, D., & Kruger, J. (2008). Why the unskilled are unaware: Further explorations of (absent) self-insight among the incompetent. *Organizational Behavior and Human Decision Processes*, 105(1), 98–121.
- Eickhoff, C., Harris, C. G., de Vries, A. P., & Srinivasan, P. (2012, August 12–16). Quality through flow and immersion: Gamifying crowdsourced relevance assessments. In *SIGIR'12. Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Portland, OR, USA (pp. 871–880). New York: ACM Press.
- Gadiraju, U., Demartini, G., Kawase, R., & Dietze, S. (2015). Human beyond the machine: Challenges and opportunities of microtask crowdsourcing. *IEEE Intelligent Systems*, 30(4), 81–85.
- Gadiraju, U., Demartini, G., Kawase, R., & Dietze, S. (2018). Crowd anatomy beyond the good and bad: Behavioral traces for crowd worker modeling and pre-selection. In *Computer Supported Cooperative Work (CSCW)* (pp. 1–27).
- Gadiraju, U., & Dietze, S. (2017). Improving learning through achievement priming in crowdsourced information finding microtasks. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference* (pp. 105–114). ACM.
- Gadiraju, U., Fetahu, B., & Kawase, R. (2015a). Training workers for improving performance in crowdsourcing microtasks. In *Proceedings of the 10th European Conference on Technology Enhanced Learning. EC-TEL 2015* (pp. 100–114). Springer.
- Gadiraju, U., Kawase, R., Dietze, S., & Demartini, G. (2015b, April 18–23). Understanding malicious behavior in crowdsourcing platforms: The case of online surveys. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, CHI 2015*, Seoul, Republic of Korea (pp. 1631–1640).
- Gadiraju, U., Fetahu, B., Kawase, R., Siehdnel, P., & Dietze, S. (2017a). Using worker self-assessments for competence-based pre-selection in crowdsourcing microtasks. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 24(4), 30.
- Gadiraju, U., Yang, J., & Bozzon, A. (2017b). Clarity is a worthwhile quality: On the role of task clarity in microtask crowdsourcing. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media* (pp. 5–14). ACM.

- Gadiraju, U., & Kawase, R. (2017). Improving reliability of crowdsourced results by detecting crowd workers with multiple identities. In *International Conference on Web Engineering* (pp. 190–205). Springer.
- Gadiraju, U., Kawase, R., & Dietze, S. (2014). A taxonomy of microtasks on the web. In *Proceedings of the 25th ACM Conference on Hypertext and Social Media* (pp. 218–223). ACM.
- Georgescu, M., Pham, D. D., Firan, C. S., Gadiraju, U., & Nejdil, W. (2014). When in doubt ask the crowd: Employing crowdsourcing for active learning. In *Proceedings of the 4th International Conference on Web Intelligence, Mining and Semantics (WIMS14)* (p. 12). ACM.
- Haas, D., Ansel, J., Gu, L., & Marcus, A. (2015). Argonaut: Macrotask crowdsourcing for complex data processing. *Proceedings of the VLDB Endowment*, 8(12), 1642–1653.
- Ho, C. J., & Vaughan, J. W. (2012). Online task assignment in crowdsourcing markets. In *AAAI* (Vol. 12, pp. 45–51).
- Howe, J. (2006). The rise of crowdsourcing. *Wired Magazine*, 14(6), 1–4.
- Irani, L. C., & Silberman, M. (2013). Turkopticon: Interrupting worker invisibility in amazon mechanical turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 611–620). ACM.
- Kaufmann, N., Schulze, T., & Veit, D. (2011, August 4–8). More than fun and money. Worker motivation in crowdsourcing—A study on mechanical turk. In *A Renaissance of Information Technology for Sustainability and Global Competitiveness. 17th Americas Conference on Information Systems, AMCIS 2011*, Detroit, Michigan, USA. Association for Information Systems.
- Kazai, G. (2011). In search of quality in crowdsourcing for search engine evaluation. In *Advances in information retrieval* (pp. 165–176). Springer.
- Kazai, G., Kamps, J., & Milic-Frayling, N. (2011). Worker types and personality traits in crowdsourcing relevance labels. In *Proceedings of the 20th ACM international conference on Information and Knowledge Management* (pp. 1941–1944). ACM.
- Kazai, G., Kamps, J., & Milic-Frayling, N. (2012). The face of quality in crowdsourcing relevance labels: Demographics, personality and labeling accuracy. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management* (pp. 2583–2586). ACM.
- Kittur, A., Chi, E. H., & Suh, B. (2008). Crowdsourcing user studies with mechanical turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 453–456). ACM.
- Kittur, A., Nickerson, J. V., Bernstein, M., Gerber, E., Shaw, A., Zimmerman, J., et al. (2013). The future of crowd work. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work* (pp. 1301–1318). ACM.
- Kosinski, M., Bachrach, Y., Kasneci, G., Van-Gael, J., & Graepel, T. (2012). Crowd IQ: Measuring the intelligence of crowdsourcing platforms. In *Proceedings of the 4th Annual ACM Web Science Conference* (pp. 151–160). ACM.
- Krajc, M., & Ortmann, A. (2008). Are the unskilled really that unaware? An alternative explanation. *Journal of Economic Psychology*, 29(5), 724–738.
- Krueger, J., & Mueller, R. A. (2002). Unskilled, unaware, or both? The better-than-average heuristic and statistical regression predict errors in estimates of own performance. *Journal of Personality and Social Psychology*, 82(2), 180.
- Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, 77(6), 1121.
- Kulkarni, C., Wei, K. P., Le, H., Chia, D., Papadopoulos, K., Cheng, J., et al. (2015). Peer and self assessment in massive online classes. In *Design Thinking Research* (pp. 131–168). Springer.
- Lykourentzou, I., Antoniou, A., Naudet, Y., & Dow, S. P. (2016). Personality matters: Balancing for personality types leads to better outcomes for crowd teams. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing* (pp. 260–273). ACM.

- Lykourantzou, I., Kraut, R. E., & Dow, S. P. (2017). Team dating leads to better online ad hoc collaborations. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, CSCW '17* (pp. 2330–2343). New York, NY, USA: ACM.
- Marshall, C. C., & Shipman, F. M. (2013). Experiences surveying the crowd: Reflections on methods, participation, and reliability. In *Proceedings of the 5th Annual ACM Web Science Conference* (pp. 234–243). ACM.
- Marston, W. M. (2013). *Emotions of normal people*. Routledge.
- Martin, D., Hanrahan, B. V., O'Neill, J., & Gupta, N. (2014). Being a turker. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing* (pp. 224–235). ACM.
- Martin, D., O'Neill, J., Gupta, N., & Hanrahan, B. V. (2016). Turking in a global labour market. *Computer Supported Cooperative Work (CSCW)*, 25(1), 39–77.
- Oleson, D., Sorokin, A., Laughlin, G. P., Hester, V., Le, J., & Biewald, L. (2011). Programmatic gold: Targeted and scalable quality assurance in crowdsourcing. *Human Computation*, 11(11).
- Pongratz, H. J. (2018). Of crowds and talents: Discursive constructions of global online labour. *New Technology, Work and Employment*, 33(1), 58–73.
- Rzeszotarski, J., & Kittur, A. (2012, October 7–10). Crowdscape: Interactively visualizing user behavior and output. In *UIST'12. Proceedings of the 25th Annual ACM Symposium on User Interface Software and Technology*, Cambridge, MA, USA (pp. 55–62). New York: ACM Press.
- Rzeszotarski, J. M., & Kittur, A. (2011, October 16–19). Instrumenting the crowd: Using implicit behavioral measures to predict task performance. In *UIST'11. Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology*, Santa Barbara, CA, USA (pp. 13–22). New York: ACM Press.
- Schmitz, H., & Lykourantzou, I. (2018). Online sequencing of non-decomposable macrotasks in expert crowdsourcing. *ACM Transactions on Social Computing*, 1(1), 1.
- Schwartz, B. (2004). *The paradox of choice: Why less is more*. New York: Ecco.
- Schwartz, B., & Ward, A. (2004). Doing better but feeling worse: The paradox of choice. In *Positive psychology in practice* (pp. 86–104).
- Sheshadri, A., & Lease, M. (2013, November 7–9). SQUARE: A benchmark for research on computing crowd consensus. In *HCOMP'13. Proceedings of the First AAAI Conference on Human Computation and Crowdsourcing*, Palm Springs, CA, USA (pp. 156–164). AAAI Press.
- Strauss, A., & Glaser, B. (1967). *Discovery of grounded theory*. Chicago: Aldine.
- Strauss, A. L. (1987). *Qualitative analysis for social scientists*. Cambridge: Cambridge University Press.
- Venanzi, M., Guiver, J., Kazai, G., Kohli, P., & Shokouhi, M. (2014, April 7–11). Community-based Bayesian aggregation models for crowdsourcing. In *WWW'14. Proceedings of the 23rd International World Wide Web Conference*, Seoul, Republic of Korea (pp. 155–164). New York: ACM Press.
- Vuurens, J. B., & De Vries, A. P. (2012). Obtaining high-quality relevance judgments using crowdsourcing. *IEEE Internet Computing*, 16(5), 20–27.
- Wang, J., Ipeiritos, P. G., & Provost, F. (2011, March 12–14). Managing crowdsourcing workers. In *WCBI'11. Proceedings of the Winter Conference on Business Intelligence*, Salt Lake City, Utah, USA (pp. 10–12). Citeseer.
- Yu, H., Shen, Z., Miao, C., & An, B. (2012, December 4–7). Challenges and opportunities for trust management in crowdsourcing. In *2012 IEEE/WIC/ACM International Conferences on Intelligent Agent Technology, IAT 2012*, Macau, China (pp. 486–493). IEEE Computer Society.

Chapter 7

Crowdsourcing and Scholarly Culture: Understanding Expertise in an Age of Populism



**Alan Dix, Rachel Cowgill, Christina Bashford, Simon McVeigh
and Rupert Ridgewell**

Abstract The increasing volume of digital material available to the humanities creates clear potential for crowdsourcing. However, tasks in the digital humanities typically do not satisfy the standard requirement for decomposition into microtasks each of which must require little expertise on behalf of the worker and little context of the broader task. Instead, humanities tasks require scholarly knowledge to perform and even where sub-tasks can be extracted, these often involve broader context of the document or corpus from which they are extracted. That is the tasks are macrotasks, resisting simple decomposition. Building on a case study from musicology, the *In Concert* project, we will explore both the barriers to crowdsourcing in the creation of digital corpora and also examples where elements of automatic processing or less-expert work are possible in a broader matrix that also includes expert microtasks and macrotasks. Crucially we will see that the macrotask–microtask distinction is nuanced: it is often possible to create a partial decomposition into less-expert microtasks with residual expert macrotasks, and crucially do this in ways that preserve scholarly values.

A. Dix (✉)
Swansea University, Swansea, UK
e-mail: alanjohndix@gmail.com

R. Cowgill
School of Music, Humanities and Media, University of Huddersfield, Huddersfield, UK
e-mail: r.e.cowgill@hud.ac.uk

C. Bashford
School of Music, University of Illinois at Urbana-Champaign, Champaign, USA
e-mail: bashford@illinois.edu

S. McVeigh
Department of Music, Goldsmiths, University of London, London, UK
e-mail: s.mcveigh@gold.ac.uk

R. Ridgewell
British Library, London, UK
e-mail: rupert.ridgewell@bl.uk

7.1 Introduction

Plato grappled with the way Socrates, his hero and mentor, had been summarily executed by the democracy of Athens; and how easy it is for democracy to slip into ochlocracy and from that to tyranny. In an age when the UK Justice Secretary could publically pronounce that “*people in this country have had enough of experts*” (Gove 2016), how do we in the academe tread the line between expertise and elitism?

In this chapter, we explore the barriers to crowdsourcing within the digital humanities. As digitised sources become ever more extensive, they overwhelm the possibility for complete analysis by traditional scholarship. Crowdsourcing and computational analysis offer ways to deal with otherwise impossible large volumes of material, and yet run the risk of simply creating voluminous trash.

Is academic resistance to crowdsourcing an elitist fear of the unwashed, or justifiable wariness of incipient poor scholarship?

We will attempt to dig into some of the core values that lie at the heart of scholarly culture, exploring how issues of authority and integrity are crucial not to the maintenance of the scholarly elite, but to the nature of scholarship itself. Through this understanding, we explore ways in which digital technology could allow wider participation whilst preserving the core values of academia.

As a case study, we draw on our experience in a particular domain: the study of the development of public musical performances through evidence of ephemera, such as notices and advertisements, and our work to create a definitive digital archive in the *In Concert* project and earlier projects.

As an academic domain, this stands in contrast to more traditional musicological approaches that place composers, performers, patrons and critics—the elite of the music world—at the centre stage. Instead, the focus on audiences, performance, ephemera and the development of print-music consumption is one that gives voice to the listener, and, to an extent, the masses.

However, taking seriously the role of mass print-culture as the subject of study does not mean these studies themselves are not expert activities. Indeed, the plethora of long-dead performers and nowadays obscure composers make the area opaque to all but the most knowledgeable. When creating a scholarly digital archive, throwing open anything but the most mundane activities to crowdsourcing appear to risk polluting the authoritative corpus.

Within the bounds of the *In Concert* project, we have not fully managed to square this circle, but we have been able to combine varying levels of expertise and automated contributions as part of a reimagined process of digital archive creation. Through this, we believe we have come closer to understanding potential ways forward, including critically the use of digital infrastructure to maintain adequate provenance to ensure that when data is viewed its authoritative, or non-authoritative status is evident. This parallels lessons from scientific crowdsourcing, which use a variety of means to develop measures of expertise, trust and degrees of certainty.

In the remainder of this chapter, we will first look at digital archives, and crucially the way the dichotomy between macrotasks and microtasks is less clear when

we consider the way expert macrotasks can be decomposed for crowdsourcing. We then proceed to describe the key case study for the chapter, the *In Concert* project, including its datasets, and some of the barriers to progress it has encountered. We consider the potential to address some of these barriers using crowdsourcing or automation, both in general within the digital humanities and considering sub-tasks within *In Concert* itself; however, we will see that crowdsourcing brings its own problems and barriers. Some of these barriers to crowdsourcing are technical, but some more fundamental, about the nature of the academic process, and so we then look at the scholarly values and academic value mechanisms that drive and constrain work in the humanities. By understanding these, the *In Concert* project was able to effectively employ automatic and non-expert human processes in various substantive sub-tasks. By studying these successful applications of non-expert, but not crowdsourced, interventions, we develop heuristics that have the potential to encourage and enable appropriate macrotask crowdsourcing in the humanities.

7.2 Crowdsourcing of Digital Archives

Crowdsourcing has already been effectively used in the digital humanities, for example in projects inviting members of the public to align historic maps with current maps.

However, it is also clear that some aspects of digital humanities are not easily amenable to crowdsourcing. Interpreting a thirteenth-century letter may require not only an understanding of the language and writing style of the time, but also an appreciation of the political and personal relationships within the court. This is evident in even relatively short time scales, for example, the mutation of the word ‘celebrity’ from the quality of a solemn occasion to a B-list reality TV star.

The case study in the chapter concerns the creation of digital archives, many dating back to just the nineteenth century, so with fewer linguistic barriers than older material, but still requiring scholarly expertise and knowledge of the time, personae and available repertoire.

7.2.1 *Corpus Creation Process*

Figure 7.1 shows a simplified view of the process for the creation of digital archives.

Stage 1 is the low-level digitisation/transcription and clearly most amenable to either automation of crowdsourcing via microtasks as, for relatively modern sources, they require little expertise beyond normal language skill set.

Stage 2 includes more complex tasks, which require more expertise. It is at this stage that the academic value of the digital corpus is largely created. The tasks even here range from those requiring deep knowledge of the period or subject matter, and some that, at first sight, may involve less expert knowledge.

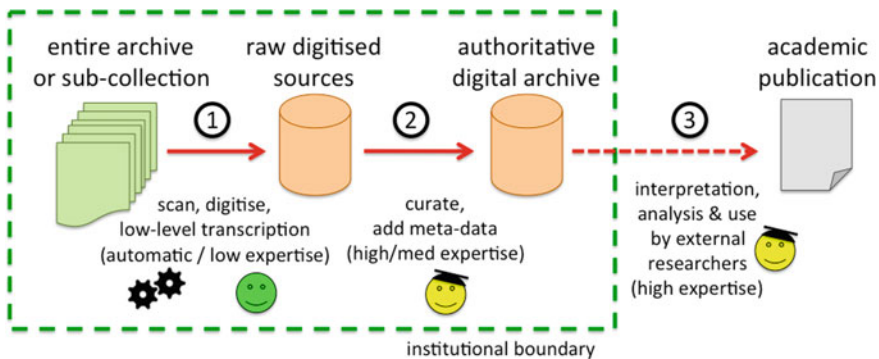


Fig. 7.1 The digital archive process (from Dix et al. 2014)

The output of this second stage is an authoritative digital archive that can be used as a base resource for further scholarship leading (stage 3) to publication: books, chapters and articles. Typically, this may first be carried out by the scholars who produced the article, but then later the authoritative archive may be released to those outside the boundaries of the original team or institution.

In this chapter we will be focusing most extensively on stages (1) and (2) and perhaps most crucially stage (2), which emerges as a bottleneck in the *In Concert* case study.

7.2.2 Macrotasks and Microtasks in Corpus Creation

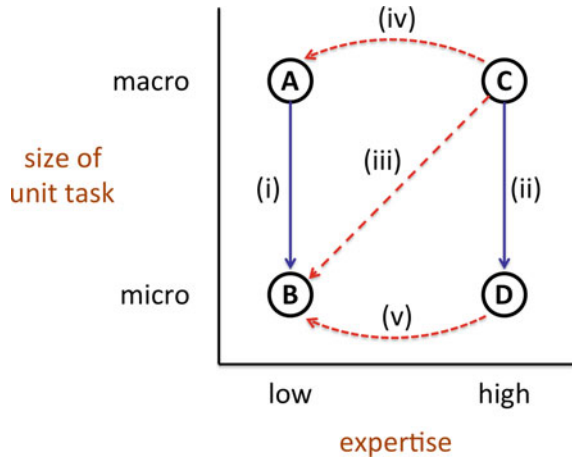
Figure 7.2 shows different kinds of the task along the axes: the *size* of individual items of the task; and the *expertise* needed to accomplish the task.

At the top left (A), we have large tasks requiring little expertise, for example, given a 1950s map and 1970 map of London align the locations of road junctions common to both. At lower left (B) we have small inexpert tasks, for example, extracting the item and cost from a single line in a receipt. At upper right (C) we have large expert tasks, for example, understanding the correspondence of a minor poet. Finally, at the lower right (D), we have small tasks requiring expertise, for example, in a single paragraph of a correspondence identifying the names of other poets of the time.

Traditional crowdsourcing is effectively about the move (i) from (A) to (B), breaking down large tasks into small parts, each able to be assigned individually to relatively inexpert workers. In contrast, traditional professional work often involves a level of task decomposition (ii) from (C) to (D); indeed this is precisely the purview of classic time-management techniques.

Of course, this is a simplification. There are many gradations of expertise, and we will see examples where there is a distinction between work that can be carried out by junior academics, and work that requires a field expert.

Fig. 7.2 Expertise and task decomposition



In the digital humanities we will typically start with large expert tasks (C), and ideally would like to break it down into many small microtasks that are amenable to low-expertise crowdsourcing (A). That is we would like to make the transition (iii).

In the simplest case, once the decomposed microtasks are performed, the overall macrotask is itself complete; for example, if we have transcribed each individual phrase of a speech, we have transcribed the whole speech. However, at very least there is a level of automatic processing, to aggregate the results of the microtasks. Furthermore, there are often residual macrotasks that need to be performed (Fig. 7.3); for example, in the map-matching task, there may be discrepancies due to crowdsourcing worker errors, complexity of the data (e.g. two streets or landmarks with the

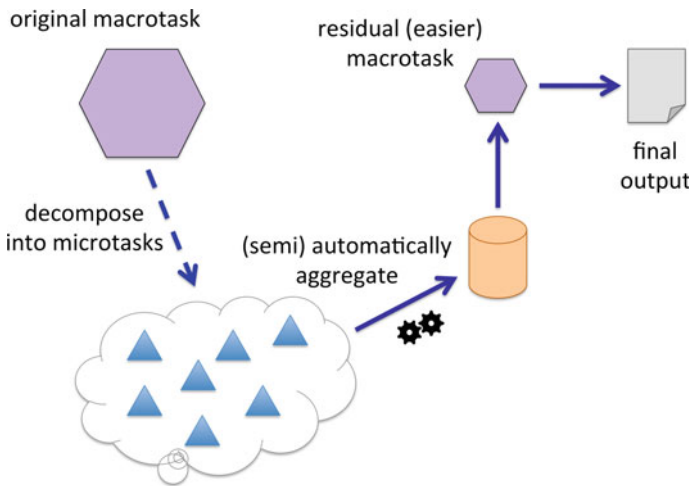


Fig. 7.3 Residual expert macrotasks

same name), or errors by the original map-maker. Often these residual macrotasks involve greater expertise than the crowdsourced microtasks, but are easier or less voluminous once the microtasks are complete. Effectively this is achieving transition (iii) by route (iv)–(i).

In fact the three stages of Fig. 7.1 can be seen as an example of transition (iii) by route (iv)–(i). The highly expert tasks of creating scholarly outputs are broken down into three stages, the first of which requires less expertise than the latter two. Furthermore, stage 1 is often amenable to decomposition into microtasks even if these are at the junior academic level rather than full crowdsourcing.

The other potential route from (C) or (B) is via route (ii)–(v). The initial expert macrotask is first decomposed into many expert microtasks and then each microtask is further decomposed into a less-expert and more-expert part (Fig. 7.4). The less-expert part may then be amenable to crowdsourcing, automatic processing or delegating to junior academics. Many of the examples we shall encounter in the *In Concert* case study fall into this pattern.

Often this microtask decomposition may be in the form of the expert microtasks that simply verify the initial less-expert microtask. There may also be some form of pre-filtering into simpler and harder macrotasks, or some form of validation that highlights discrepancies, or other cases requiring more expert interventions.

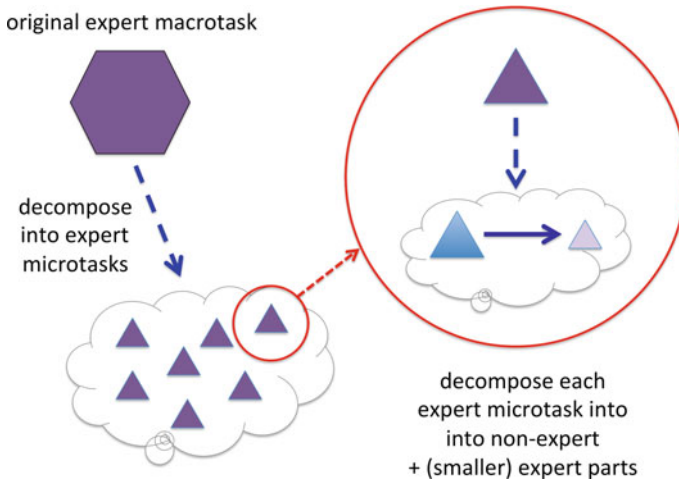


Fig. 7.4 Decomposing microtasks

7.2.3 *The Myth of the Acontextual*

Haas et al. (2015) describe a microtask in terms of questions that “require little context or training to answer”. We have discussed the ‘little training’ aspect in terms of expertise, but the acontextual element also requires examination.

Clearly, some tasks require an understanding of a whole corpus, for example, assessing the mood of a particular politician from reading correspondence written during the lead up to a critical event. There are crowdsourcing techniques targeted at such non-decomposable tasks. Notably, TAS (Task Assignment and Sequencing) passes large tasks of this kind sequentially between a number of crowd workers; each spends considerable time on the task as a whole, advancing work on it, before passing on to another (Schmitz and Lykourantzou 2018). An R&D task was used for the empirical evaluation of TAS, which also included a pre-test for knowledge of the domain (e.g. FIFA); that is an element of crowdworker expertise.

In many tasks, the decomposable/non-decomposable distinction is less dichotomous than first appears. Think of the map-matching tasks. Maps may vary in the way they portray different features, for example, showing built-up areas as blocks of colour or divided into individual properties; or they may use different abbreviations. Although each atomic matching task is relatively independent, still there will be a level of learning as the task is performed.

Sometimes, the microtasks only make sense within the larger context, for example, we will see in Fig. 7.9, how one of the musicologists in *In Concert* spreads out paper across a table as part of what appears to be a more focused matching task.

Furthermore, some of this learning is likely to feed into higher level understanding. Spending time identifying key names and events from a politician's letters may seem like a low-level task, but of course, is immersing the reader in the life of the writer. Indeed, one method for dealing with creative tasks is to deliberately create a ‘busy work’ aspect, which can be performed fairly automatically but is at the same time orienting one's mind towards the larger creative task (Dix 2019).

Any outsourcing of microtasks to crowd work or automation needs to be cognizant of these subtle, but crucial effects (Fig. 7.5). For example, very early CAD systems were introduced in architects' offices in the late 1970s in order to reduce the time-consuming tracing of plans from previous projects, which was often the first stage in starting a new related project. Although it certainly sped up the drafting process, the architects found themselves more highly stressed and less productive overall: the low-level tracing activity had been giving them precious time to think about and prepare for the creative task.

7.3 *The In Concert Project*

This chapter draws on case studies from *In Concert: Towards a Collaborative Digital Archive of Musical Ephemera* (2014–2016), a sub-project of the AHRC funded *Trans-*

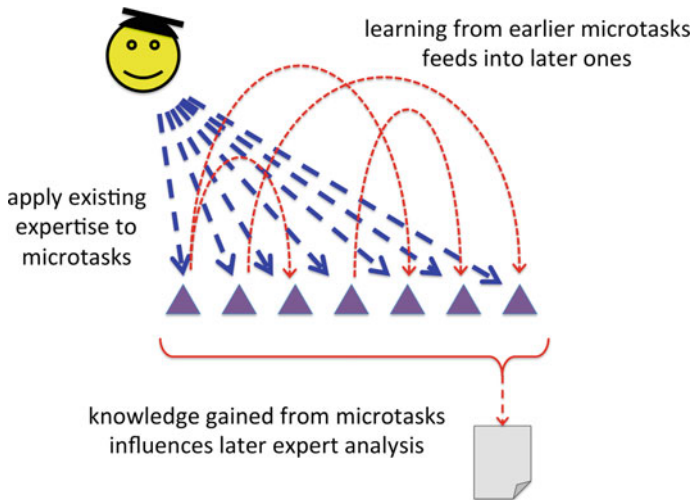


Fig. 7.5 Microtasks lead to understanding

forming Musicology programme (2016). This project was a collaboration between Musicology and Computer Science and had a dual aim. On the one hand, the musicology goal was to enhance a number of datasets related to concerts in London from the eighteenth century onwards. However, there was also a broader digital humanities goal to use this experience to better understand the evolving role of the digital archive. Indeed, this chapter is one of the outcomes of this broader goal.

7.3.1 *Performance and Ephemera*

Much of musicology is focused on composers and their works. This may include the historical study of the lives of the composers and of the development of individual works from sources such as letters, contemporary accounts, and official records. In the way of academia, this involves a highly specialised study of relatively sparse sources.

In contrast *In Concert* was focused on the actual performance of music—what was selected and listened to rather than what was produced. The canon, the works that persist, is not merely about the ‘genius’ of the individual composer, but also the trends within the broader culture. Furthermore, the patterns and trends of performance and performer are not merely reflections of taste, but also connect to issues in social history such as the role of women and the privileging of repertoire that reflects the interests and identity of the culturally empowered (Cowgill and Poriss 2012).

Today, the consumption of music may be studied directly and near instantaneously through streaming services such as Spotify, downloads from iTunes, the schedules of BBC Radio, or even, for popular music, the long-running weekly ‘top 20’. However,

in pre-internet days, the sources are more diverse and dispersed, often in the form of *ephemera*: concert programmes, newspaper reports and advertisements; things never intended to be preserved for posterity (although historically concert programmes were often collected and bound).

As with any historical source, there is partiality and bias in what was reported and what was preserved: the concerts of high society are more visible than the songs sung in taverns. However, to give the most reliable picture of the patterns of performance, the ephemera needs to be sampled, collated and recorded in a consistent and scholarly manner.

7.3.2 Datasets

In Concert focused on three principal datasets:

- LC18—*Calendar of London Concerts 1750–1800* (McVeigh 1992–2014)—This was created from many sources relating to concerts in the second half of the eighteenth century. Given the relatively small number of sources and events during the period it is a near-exhaustive collection of available information.
- LC19—*Concert Life in Nineteenth-Century London* (Bashford et al. 2000)—By the nineteenth century, the number of concerts and relevant print sources grew to such an extent that a complete compilation is not possible; instead, sample years at 20-year intervals were exhaustively studied, using newspaper archives and other sources.
- CPE—*Concert Programme Exchange (Konzertprogramm Austausch) 1901–1914*—In the early years of the twentieth-century Leipzig publisher Breitkopf & Härtel distributed printed copies of programmes of major concert venues in Europe, Russia and America. The British Library’s collection of these was digitised for Gale Cengage, making around 12,000 programmes available in OCR form.

These were supplemented with two other datasets primarily as authority files:

- CPP—*Concert Programmes Project (2004–2007)*—This project, administered at the British Library, collates meta-information about archives; it does not contain programmes or programme text itself, but lists archives and collections (most offline) where such ephemera can be found and information about the venues and people they cover.
- BMB—*British Musical Biography 1897* (Brown and Stratton 1897)—This 400-page volume includes nearly 4000 entries for British musicians and composers during the nineteenth century and is broadly contemporary with the first editions of Grove (1900). A digital version was created as part of the *In Concert* project based on OCR from the Internet Archive.

These data sources overlap in terms of subject, people and venues, but represent very different stages of digitisation from raw OCR (CPE, BMB) to fully authoritative

corpus (LC18) and from full details of individual concert programmes (CPE) to meta-information about the presence of archives (CPP).

7.3.3 *Barriers to Progress*

One of the drivers for the *In Concert* project was a hiatus in the development of the LC19 dataset. As noted the sources for the nineteenth century are far more extensive than used for LC18. The 1750–1800 dataset LC18 had been the work of an individual, whereas LC19 required a team project including three primary investigators and a substantial number of research assistants.

Funded projects (1997–2001) in the mid-late 1990s were used for a first phase of the LC19 development. A relational database structure was created based on the experience of LC18 and this was initially populated by the research assistants extracting information from primary sources, principally newspaper adverts and notices and concert programmes. The research assistants would create a record for each advert/notice and fill in details such as the date, venue, performers, works and composers in the programme. This was successfully completed and this first phase data was used as the basis of initial analysis and publications (Bashford et al. 2000).

However, a second phase was always envisaged. The initial data collection was effectively ‘raw’ data: one entry per notice, and raw text names of people and venues. The plan was to create a dataset with a single entry per concert, critically editing partial information from multiple notices, linking people, venues, works, etc. to unique authority identifiers (e.g. if there were several variants spelling of the same person’s name, or several distinct people shared a common name).

This second phase would have allowed better connections with the LC18 dataset and also statistical analysis of historical trends, visualisations, etc. In particular, LC19 has authority files (people, venues, works), which could be connected to the authority files in LC18.

However, this second ‘interpretative’ phase required more expertise and professional judgement, and so needed the time of the more senior academics, which of course is limited. Consequentially progress on this second phase had stalled for some time.

7.4 Opportunities for Crowdsourcing and Automation

7.4.1 *Challenges of Scale*

The difference between the LC18 dataset and LC19 demonstrate the challenges of scale inherent in digital humanities research. LC18, with about 4000 concerts, was already an extensive exercise, but was possible by a single dedicated scholar. How-

ever, the increase in the number of popular performances and available sources in the nineteenth century meant that even a 1-year-in-twenty sample required a substantial team effort and its final phase was incomplete. Even this belies the fact that the volume was changing throughout the nineteenth century, so that, by the time we come to the twentieth century, archives of individual concert houses are themselves of similar or greater extent and CPP's meta-records of these archives are themselves large.

This increase in volume is a result partly of a greater number of events, but also the greater number of preserved sources, the problem of the 'infinite archive' (Bell 2004). For classicists or traditional scholars dealing with sparse sources, the problem is interpreting the little data that is available. For 'born digital' materials, such as Spotify logs, the issues are almost those highlighted by Borges' imaginary map (Borges 1946) where the data is almost coincident with the world itself; given the massive volume data the problem is what to select or even ignore in order to turn raw data into information.

Between these extremes are areas such as those dealt with in *In Concert*, where the number of physical or raw digitised resources is too great for scholars ever to deal with, and yet requires a level of processing and interpretation before it is suitable for scholarly analysis.

Note this does not invalidate traditional scholarly approaches to historical archives. If you have a focused topic of study such as the works of a minor composer, or performances in a particular venue, you still need to trawl through multiple archives to find heterogeneous sources. Although even raw digitisation may make this easier allowing faster searching and less travel to view originals if not necessary, and certainly avoiding speculative journeys only to find there are no useful resources.

However, it is a problem for the scholar wishing to study broader questions such as different patterns of repertoire between European centres, or the changes in musical taste in London venues during the nineteenth century.

Sampling, as in the LC19 dataset, is a partial way to deal with this issue, but, as we have seen, even a high level of sampling can still lead to datasets too large for expert scholarly curation.

This impasse appears to create an obvious opportunity for crowdsourcing or automated solutions.

7.4.2 *Candidate Tasks*

Looking through the various datasets we can identify a variety of tasks requiring different levels of expertise and hence potentially offering opportunities.

T1 Low-level digitisation. This may require special equipment for high-quality photography or scanning, but also may include tasks such as transcription, or correction of OCR. For example, in CPE the title pages of concert programmes often use decorative scripts, which are hard to OCR.

- T2 Identification of format or general language features. For example, in BMB the transcription included page headers, capitalised entry names, etc. In CPE concert programmes often included columns of names.
- T3 Identification, marking or extraction of semantic fields. For example, in LC19 finding the name of the venue in a newspaper advert. Another example in LC19 was the initial identification that a portion of a newspaper page was, in fact, a concert notice.
- T4 Matching text names of venues, people and works to unique entries in authority files. For example, there may be two John Smiths, father and son, who can be disambiguated by the date of the concert or the style of music. This task might also require knowing that certain performers had multiple stage names, or that a venue changed name.
- T5 Matching authority files between datasets. The LC18, LC19, BMB, and CPP all include unique identifiers of performers and composers and LC18, LC19 and CPP include venue identifiers. By connecting these not only is it possible to analyse the datasets together, but also where one dataset includes information such as external identifiers or geocoding, this becomes shared by the other datasets reducing work.
- T6 Grouping notices (in LC19) that refer to the same concert. In some cases this may simply be that two notices refer to the same venue on the same date, but some venues are large enough to have several concerts on a single day, also some notices may be vague about times, may have errors, or dates may change if a concert is postponed. In short, even the most simple concert notice/programme often has rich many-to-one relational complexity.
- T7 Merging groups of notices into a single definitive concert record. In some cases, this may simply be filling in details that are missed in one notice with complementary information in another. However, on other occasions, this may require choices between conflicting information.
- T8 Musicological analysis of the dataset. This may be by hand or by using data processing, statistical, or visualisation techniques.

Looking back to Fig. 7.1, tasks T1–T3 belong roughly to stage 1, T4–T7 to stage 2 and T8 corresponds to stage 3. It is clear that some of these tasks require less musicological expertise than others. In LC19's first phase the research assistants performed T1 and T3 (and T2 where relevant) but T4, T6 and T7 were left for more expert processing in phase 2.

7.4.3 Barriers to Crowdsourcing—Low-Level

Transcription or correction of OCR sound like straightforward candidate microtasks for crowdsourcing. However, it is interesting that the raw OCR text of BMB, (and similar documents) at the Internet Archive appears uncorrected. This appears to be partly related to complexity. It is possible for readers to correct OCR and then upload

corrected versions, but this really requires a volunteer to commit to correcting all, or a substantial part of a volume. This complexity barrier has been partly addressed by other projects.

reCAPTCHA was originally used on the New York Times archive, and it was proposed in some reports that it could be used for the Internet Archive (von Ahn et al. 2008), but it is not clear whether this ever occurred before reCAPTCHA was acquired by Google.

Distributed Proofreaders (2018) is a web-based service set up originally to help volunteer correction of Project Gutenberg texts. It allows page-by-page correction and manages different stages of proof correction from first OCR scans to more complex verification. However scanning the title of volumes processed, it is evident that the majority are either novels or books of a largely textual nature [e.g. Ackerman (1922)]. Tomes such as the British Musical Biography or gazetteers are less obvious candidates for the volunteer.

Historical texts are also harder to OCR (less distinct fonts, poorer quality paper and printing, non-standard spellings). In the case of concert programmes and notices, a great deal of information is also communicated via changes of font and tabular positioning on the page, similarly catalogue-style books such as directories, dictionaries and gazetteers often include abbreviations and special conventions, some of which, such as bolding, may be difficult to retain in OCR. There have been projects to create special-purpose OCR tools and workflows for historical texts, for example, the PoCoTo open-source software (Vobl et al. 2014) and Fink et al.'s system to create adaptive OCR based on previous proof corrections (Fink et al. 2017). However, to date, these are not part of the Internet Archive's standard workflow.

7.4.4 *Barriers to Crowdsourcing—More Complex Tasks*

As noted, the hiatus in the LC19 dataset was at a stage way beyond these low-level tasks. Academically trained research assistants read physical or digital copies of newspapers, found references to concerts and then extracted all available relevant information to input into the SQL database. This was already deemed a task requiring a level of academic expertise and training to use the database, although some aspects of the tasks might well have been possible to crowdsource (e.g. locating concert notices).

However, even the research assistants were not deemed sufficiently expert to perform tasks T4, T6 and T7 on the LC19 dataset. To an outsider aspects of these tasks look as though they could be suitable for crowdsourcing. For example, T6, grouping multiple notices that relate to the same concert, appears to be something that is possible based on general knowledge and understanding: looking through date ordered lists of notices, and collecting those that appear to be at the same venue at the same time.

Early in *In Concert*, the potential for using knowledgeable amateurs for crowdsourcing was discussed. These were often referred to as 'Radio 3 listeners'—Radio 3

is the BBC classical music radio channel in the UK, and listeners tend to come from both a slightly older and more highly educated demographic than the general population. The general idea of using such knowledgeable crowdsourcing was accepted as a good idea, but any suggestion of actually doing this for specific tasks in the dataset was greeted with concern, the idea of the musical amateur seeming to be at odds with that of the scholarly corpus.

Again, looking from the outside, this at first may seem to be a case of excessive scholarly purity. However, digging deeper it relates to justifiable caution—any uses of crowdsourcing for macrotasks needs to be done in ways that understand and fit within the overall scholarly culture. We should note that we were not the first music-based project to have to deal with problems in this area (Bodleian Library [2012/2019](#)).

7.5 Scholarly Values and Academic Value

Key to the success of any system deployment whether digital, physical or organisational, is an understanding of the underlying values and value within the setting.

- *Individual values*—What are the internal beliefs, motivations and drivers that create a sense of personal worth and lead individuals or groups to judge the worthwhile nature of outcomes?
- *Value mechanisms*—What are the external measures, rewards and validation offered by the wider system in which individuals or groups participate?

Those entering academia on the whole assent to a number of common scholarly values such as integrity and the desire to increase the bounds of scholarship. However, they also operate within a matrix of reward and career advancement mechanisms including promotion procedures, metrics for external assessment (such as the UK REF), and publication routes.

In previous work, we have explored the values and value mechanisms that are critical in forming attitudes towards crowdsourcing and automation within digital humanities (Dix et al. [2014](#)). We will summarise these as a basis for understanding potential ways forward.

7.5.1 *Scholarly Values: Authoritative and Complete*

The term *authoritative* in the above is crucial both for the scholar's own use and for the scholar to be happy for others to see the work. The methods of creation need to be well-documented and of consistently high quality so that further scholarship can be built upon it.

In some cases, the corpus may not be exhaustive, but it is important that it is complete in the sense of covering a known period, geographic area or other selected

(and stated) criteria. This may include sampling, as has been done with LC19, but in this case, the sample needs to be unbiased and clear in its criteria.

In essence, this is about the ability of the scholar using the corpus to be able to assess the reliability of data within it and make defensible inferences and arguments based upon it.

Any dataset inevitably embodies potential bias in the collection and preservation methods (as noted history selects for the rich and powerful) and also in interpretation. Indeed scholars differ in their approaches to the record and each scholar's use of a resource will vary depending on their assessment of the curator's hermeneutic.

In this context, a distrust of the amateur is understandable. If a known scholar has curated a digital archive, then those using it can take this into account; even if they disagree with the curator, they can still rely on basic levels of scholarly consistency and accuracy. If many amateur hands are at work during crowdsourcing it seems impossible to know if all of the data is of sufficient quality without checking everything, and furthermore, different workers may make inconsistent decisions.

As well as potential problems in the use of the resulting corpus, those in charge of curating the digital archive feel responsible for it. If there are inaccuracies or omissions, they will feel they are letting down their own personal standards and potentially weakening their academic credibility and reputation.

7.5.2 Academic Reward

Stage 3 of Fig. 7.1 includes the digital archive being available to the wider research body as well as the scholars involved in its creation. In practice, this may be delayed for many years, or even indefinitely.

One reason is related to the scholarly values above: the curator(s) need to be very sure they are releasing a corpus on which they feel comfortable to rest their scholarly reputation. Preparing a corpus to the point where you can perform your own research is less onerous as you understand the limitations and sources of various parts, and so are able to make assessments of validity.

Intellectual property issues are also problematic: some sources restrict access to personal research, one's rights to republish derived datasets may be unclear, and it may be hard to determine the correct licence under which to release one's own data.

Technical barriers may also deter publication of data. Although this is becoming easier as many universities create digital repositories, the complexity and costs of digital archiving are perhaps underlined by the UK Arts and Humanities Research Council (AHRC). During the 2000s the AHRC mandated that all funded projects lodge their resulting data in the AHRC's own archive, the Arts and Humanities Data Service, possibly in the process leading researchers to believe this was an archival store backed by the resources of government. However, by the end of the decade, the AHRC not only dropped the requirement, but closed the repository (Chris Rusbridge 2007; Wikipedia 2019).

This story underlines the ambiguous role of data in the research process and hence the most critical reason for delaying dataset publication.

Broad scholarly values lead one towards openness, expanding the breadth of knowledge. However academic reward mechanisms both formal and informal are oriented primarily towards scholarly publication in books or journal articles (depending on the discipline). Although the community will be grateful to the scholar who makes curated resources available, the real academic applaud goes to the scholars who interpret those resources and create publications from them.

In the UK Research Excellence Framework, the periodic assessment of national academic research, Panel D, which covers arts and humanities, did include a curated ‘database’ as a valid research output (Research Excellence Framework 2012). However, Panel B (science and engineering) did not mention data as a valid output at all, despite the Web being developed by Berners Lee precisely to share scientific data from CERN (Tim Berners-Lee 1989). The Leverhulme Trust, which funds cross-disciplinary research is even more specific explicitly rejecting applications where “*the balance between assembling a data bank or database and the related subsequent research is heavily inclined to the former*” (Leverhulme Trust 2018).

In summary, academia regards the publication of data as *valuable*, but does not *value* it.

7.6 Radical Transformations to Support Traditional Values

Having disentangled some of the complex web of values and reward mechanisms that underlay the scholarly curation process, our challenge within *In Concert* was to radically reimagine that process in ways that preserve the underlying scholarly values and work within the academic reward mechanisms and yet are more open in terms of both publication of data and accepting automated or non-expert input.

As noted earlier in this chapter, we did not adopt crowdsourcing. This was partly for reasons of time, and partly because the team was still resolving the issues and barriers as described. However, we did use automated algorithms, which on the surface have some similar problems to crowdsourced work, and also two human non-experts (a very small crowd), the technical partner in the project and another non-expert known to the team.

We will describe three tasks in the project where these non-experts (human and machine) formed part of the process and then return to reflect on the lessons this has for future crowdsourcing of macrotasks in the humanities. Each of these follows broadly the route (ii)–(v) outlined in Sect. 7.2.2. Each takes an initial macrotask, creates a combination of non-expert microtasks, semi-independent expert microtasks, and residual expert macrotasks.

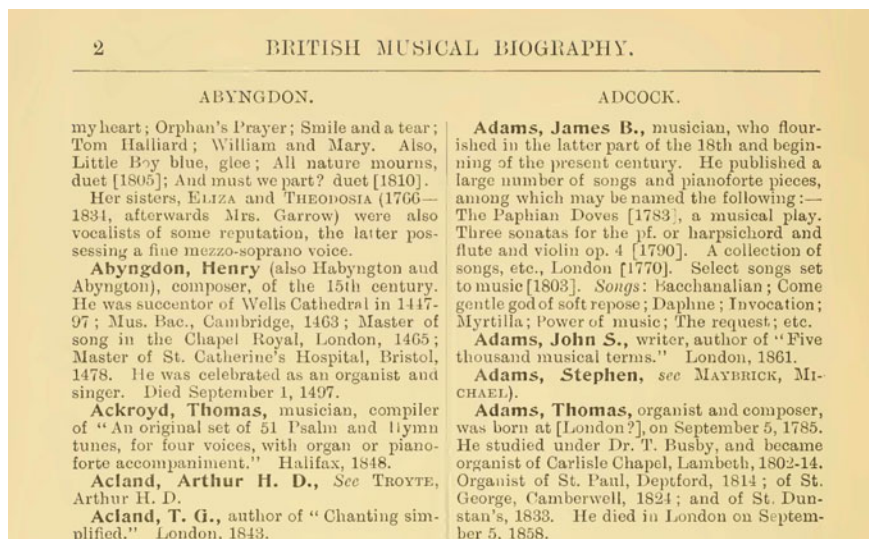


Fig. 7.6 Portion of Brown and Stratton's *British Musical Biography* (Brown and Stratton 1897)

7.6.1 A Digital Version of the *British Musical Biography*

One of the core datasets of *In Concert* was CPE, the *Concert Programme Exchange*, which was at the earliest stage of preparation with OCR only. The range and complexity of the documents, concert programmes from many venues, meant that further automatic processing would be very difficult. It is an ideal candidate for both low-level crowdsourcing, tidying up OCR of florid fonts, and also higher level tasks such as marking up titles, players of different instruments, pieces performed, etc.

The *British Musical Biography* (BMB) was at a similar stage of preparation, with a raw OCR at the Internet Archive, but its strong structure made it far more amenable to automated processing (see Fig. 7.6). This was valuable in its own right, but, more important, acted as an exemplar allowing the project to learn lessons and develop processes which, we hope, would be useful for the more complex CPE.

Page breaks were not marked in the OCR, but the page number and capitalised 'BRITISH MUSICAL BIOGRAPHY' were a (relatively) easy marker for automated pagination, similarly the capitalised column headings made them (relatively) easy to spot automatically. The bold font was not marked in the OCR, but the entries are of the form:

Name, Name {optional initials}

where the names have initial capitals and there are a small number of variations. This allowed the entries to also be identified.

These automatic structuring rules were supplemented with sanity checking rules, for example, verifying that page numbers are consecutive and entry names in alphabetical order.

If the original text and OCR had been perfect, this would have enabled computational algorithms to process the text unaided. However, this was not the case. The quality of the print led to frequent OCR errors, for example, some capitals (such as ‘C’) could be read as lower case, lower case L as a bar ‘l’, and commas and full stops could be confused. Added to this there were some errors in the text itself such as comma/full-stop mistakes in typesetting and names out of proper alphabetic order. Finally, although most names fell into simple patterns, others, for example, royalty, required specialised rules.

Where failures in sanity checks were attributable to incorrect OCR, the OCR text was edited by hand and the files re-processed. Other failures led to refinements of the rules, for example, different name formats. However, in some cases *exception files* were created, that is tables of specific rules such as: “the entry on line 27 of the right hand column on page 23 should read Doe, John”. These exception files have become a recurrent pattern in our attempts to automate different forms of processing: not everything can be captured in generic rules.

Finally, a page-by-page check was made to verify that the database entries did correspond to those in the OCR, although there was no attempt to completely fix the OCR in the text within an entry.

It should be noted that the hand checking was carried out completely by the non-experts, and would almost certainly have been possible as a crowd-sourced exercise.

There are a number of factors that made this a possible task for non-experts:

1. The authoritative nature of the work was actually carried out by Brown and Stratton in the nineteenth century, this exercise was merely a digitising of an existing scholarly resource. Although the kinds of checks and rule creation varied in complexity, there were, therefore, no scholarly judgements required.
2. Furthermore, because this was not the musicologists’ own scholarly work, and merely a digitisation exercise, there was little risk of the work reflecting badly on the scholars’ reputations.
3. The non-experts were known by the team and trusted to be meticulous, for example not correcting apparent misspellings in the text as printed, merely ensuring that the digitised form corresponded to the page.

7.6.2 Cross-Linking Authority Files

We had name and place information from four sources. Both the 1750–1800 and 19th Century London Concert datasets (LC18 & LC19) have authority files for people (composers and performers) and places (venues). The Concert Programmes Project (CPP) has large authority files for places and agents (people, groups and organi-

sations), including some geo-referencing and planned VIAF links. British Musical Biography (BMB) has people's names only, but is comprehensive.

Automatic matching was used to create candidate matches followed by a hand verification stage. The latter was crucial as the authoritative nature of the data was a key academic value for the humanities researchers (Dix et al. 2014); automatic matching, whilst useful, is bound to be inaccurate, yielding both false positives and false negatives. Following the principles of 'appropriate intelligence' (Dix et al. 2000), the automatic algorithms were not designed to be as clever as possible, but instead to be part of a human-computer system that as a whole yields reliable results.

7.6.2.1 Automatic Matching

Places were simplest to match automatically using plain word matching and permuted word indexes for efficiency. There are fewer place names than people's names and they tended to be more standardised; so simple matching was sufficient for candidate identification.

People names were more complex. First, this was because the data sources needed an element of cleaning/normalisation. In the LC18 dataset, the ids included an encoding of the surname, gender and possible disambiguation; for example "KNEISEL~" for the female (trailing tilde) "Henriette Kneisel", or "TURNER-2" for one of two "Turner"s. This was relatively straightforward pattern matching. More complex was the CPP data, which included groups and organisations as well as people and also was itself garnered from multiple sources. Some people's names had the forename as a separate field, some were in 'surname, first name' format, and some were more complex, including honorifics. In the spirit of maintaining the original source as 'golden copy', this task was managed through a combination of keywords for terms in organisations (e.g. 'orchestra', 'Staatstheater'), extensive lists of honorifics (e.g. 'Prince', 'Mlle', 'Duke of'), and explicit exceptions (e.g. that record id '2173' named 'Tate Britain' is an organisation not someone with surname 'Britain').

Having normalised names as much as possible, the automatic algorithm matched between datasets using a similar word match measure to the places. Fuzzy matches were not used, as this led to too many false positives and the point of the algorithm was to aid not replace human matching. Note that while crude whole word matching was used for the batch processing for names, fast fuzzy search is enabled in online datasets using both Soundex and 'drop one character' indexes. The latter stores every combination of each name with single characters dropped; by doing the same for retrieval terms one can obtain a good triage pass before more sophisticated edit distance measures are calculated.

7.6.2.2 Human Processing

Having obtained automatic 'candidate matches', these were then available for human verification via two interfaces. In one the match lists were exported as a spreadsheet

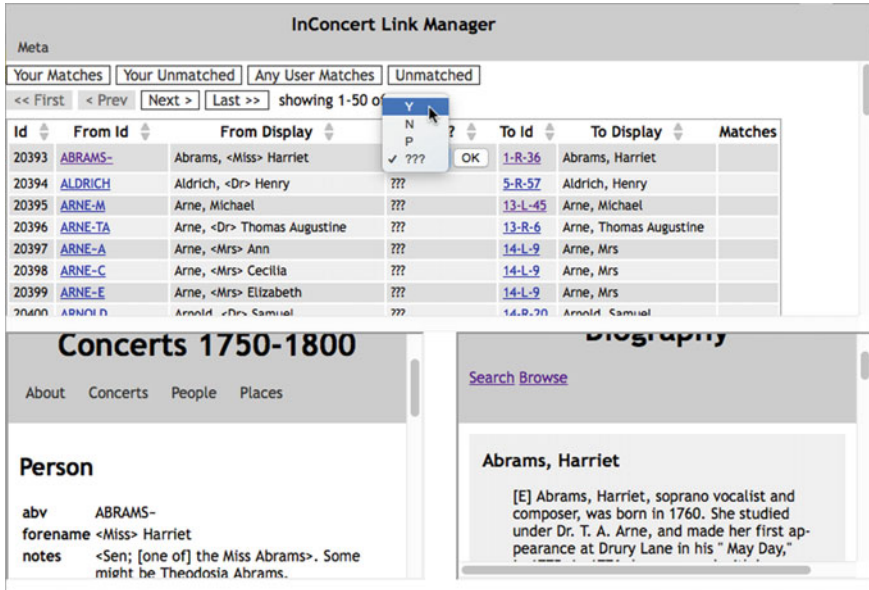


Fig. 7.7 Prototype web interface for link checking

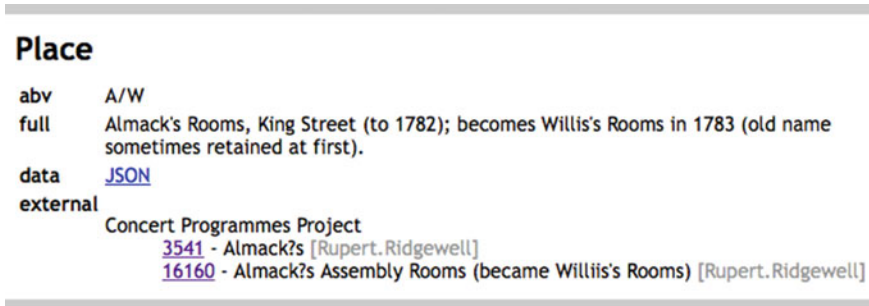


Fig. 7.8 Links displayed with provenance

for offline processing, which could then be later re-imported; in the other (Fig. 7.7), the data was presented in a web interface. Both were showing names from one data set (the source) on the left, the possible matches (targets) on the right, and a computer-generated confidence value between. The musicologist could then mark these as 'Y' (yes), 'N' (no) or 'P' (not sure).

In general, verified matches were almost always for the entry with highest automatic confidence score; however, there was no sensible 'critical value' for this confidence score, highlighting the need for human expert evaluation.

The completed spreadsheet or web interaction was processed to create a linked dataset listing the connections between the datasets (similar to RDF 'sameAs').

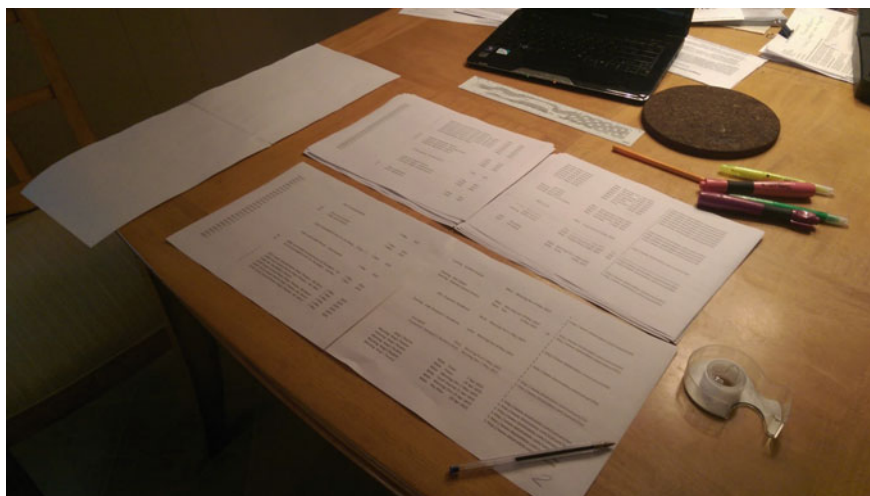


Fig. 7.9 Printed spreadsheet for grouping by hand

By keeping this separate, it is possible to easily maintain the provenance of the link information, fully automatic or human, and if human by whom (see Fig. 7.8). Different experts may resolve the names in different ways, or decide whether they trust the source of the linkage information (automatic or human) for a particular scholarly purpose. This cross-linking was also used to enable RDF Linked-Data views of the datasets (Fuller et al. 2016) (Fig. 7.9).

Note that while some of this matching was done by the musicologist, some was also performed by the technology partner, who was not an expert. However, the fact that the linking dataset contained provenance: who or what did the matching, made it possible to regard the non-expert's matching as a suggestion, just like the automatic matching. Furthermore, the ability to visualise this provenance (as in Fig. 7.2), means that anyone wishing to make scholarly judgements based on the dataset can take into account the expertise of the matcher.

7.6.3 *Grouping and Matching Within a Dataset*

As noted previously, the LC19 dataset of nineteenth-century concert notices/adverts could potentially contain multiple entries relating to the same concert. The remaining (interpretation) phase was to go through these concert notices, work out which ones referred to the same event and create an authoritative entry for each concert. This process the musicologists refer to as 'skewering', but database technologists would think of as entity/object identification or record linkage (Dunn 1946; Ahmed et al. 2007).

This process had acted as a block to progress, as it was so substantial and required expert attention. A major breakthrough was realising that this consisted of (at least) two separable sub-tasks, described earlier: (T6) match—‘skewer’ multiple notices referring to the same concert; (T7) merge—combine the data from the notices to create an authoritative record for the concert. It became clear that, while the effort in doing the match task was substantially less than the merge task, still the dataset would become substantially more valuable once the first sub-task was complete.

There is a substantial literature on entity/object identification dating back from the early days of databases (Dunn 1946) to semantic web applications (Nikolov et al. 2012). Sometimes this involves simple similarity measures such as Jacquard distance between feature sets, or Levenshtein edit distance for string matching. Other researchers have used complex machine learning techniques, including using structural relationships in relational or graph databases (Rendle and Schmidt-Thieme 2006; Bhattacharya and Getoor 2007; Di Gioia et al. 2010). There is also tool support. OpenRefine (formerly Google Refine) supports the management of data including linking names to entities (possibly more like the name matching in the previous section), although it does not do matching itself, passing this task on to external data services through its Reconciliation Service API (OpenRefine 2018). RELAIS (REcord Linkage At IStat) is dedicated to the process of record linkage itself (Scanapiego et al. 2015); it supports a number of different matching algorithms that can be applied to any combination of fields.

However, as with the name matching, because this was part of human–computer process, simpler automatic matching was sufficient combined with methods to make the human task easier. Crucially the matching algorithm was *liberal* in terms of finding potential matches: those that had the same date and similar venue names were matched into groups. This inevitably led to some false negatives (e.g. if the date or venue of a concert changed between notices) and false positives (several concerts at the same venue on the same day). However, the liberal matching was combined with a *conservative* process of marking warnings on those where the match was not almost exact.

This combination meant that it was highly likely that potential matches were already grouped, even if some groups contained more than one event. However, the warnings helped to focus attention on groups which might need division by the expert.

A similar process of exporting and importing spreadsheets was used as for the authority file matching, a process that we found extremely efficient in terms of both development time and ease of learning (Dix et al. 2016). As with the previous interface, the spreadsheet allowed the assessor to attach a level of confidence to the grouping and when the spreadsheet was re-imported, the dataset was updated to include who had performed the group verification.

7.7 Discussion—The Future for Crowdsourcing in Digital Archives

We saw that while there appear to be many potential tasks suitable for crowdsourcing when preparing a digital archive in the humanities, there are also barriers, especially for macrotasks, which tend to require a level of expertise. However, we have also seen that *In Concert* has employed both automated algorithms and (trusted) non-experts when working on the creation of its datasets.

Based on these experiences, we can revisit the issue of crowdsourcing, looking at the properties of tasks, interfaces and workflows that made it possible to use these non-expert actors. Doing this we see ways in which crowdsourcing by amateurs may be possible within a scholarly culture and identify enabling heuristics.

1. *Understanding values*—Our first and most important step was to understand the scholarly values and academic value systems that drive and constrain scholarly activity. Attitudes that, to an outsider, might seem like academic elitism are in fact rooted in the very real need to maintain a reliable and authoritative corpus. Specific practices may be radically reimagined, but only by understanding and working within a context of deep scholarly values.
2. *Deconstructing tasks*—Macrotasks where scholarly expertise seems essential may be broken down into microtasks, some of which may be amenable to less expert help, with residual less-extensive expert macrotasks. Within *In Concert* this was highly effective in recognising opportunities for automatic processing, but the same process could identify crowdsourcing potential. Crucially, there is evidence that, where it is possible, decomposing into microtasks increase the quality of results (Cheng et al. 2015) which fits well with the scholarly values. Furthermore, the lower volume of the residual macrotasks may make it easier for the scholar to apply contextual understanding.
3. *Deconstructing expertise*—Computer processing may lead to erroneous, weird and occasionally risible outputs, but it is consistent. The trusted non-experts lacked domain knowledge, but were meticulous and (in general terms) scholarly in their approach. Microtask crowdsourcing makes use of very generic low-level skills, such as visual matching. Macrotasks may involve more complex activities, for example, scanning sources for mentions of concerts, but not necessarily the knowledge of the professoriate. Distinguishing types of expertise and skill may help identify places where the ‘expert’ need not be a domain expert.
4. *Sanity check rules*—These build confidence in the processed data, but also highlight where more expert human intervention is required. In the automated processing this led to updating of rules, or creation of exceptions. In crowdsourced processing, this might lead to updating instructions or marking of certain parts of the dataset for more expert processing. Furthermore sanity checking itself may be human activity, for example the OCR correction workflow in Distributed Proofreaders (2018) involves multiple human checking stages.
5. *Suggest/confirm workflows*—In both authority file matching and concert notice grouping, the automated matching was seen as creating suggestions for expert

confirmation. In the end, the experts verified every decision, but for some kinds of tasks scanning work and confirming it can be much faster than doing the task in the first place. Crucially, this means that the expert retains control over the final output.

6. *Provenance*—Tracking provenance (who did what and identifying original sources), is, of course, essential for suggest/confirm workflows, but potentially offers the ability to have datasets with mixed levels of authority. For traditional scholarly work, where the scholar examines individual sources, they can make a case-by-case assessment of the extent to which they trust judgements by different individuals in the creation of a digital record. In some cases, if they are uncertain, they can, of course, check the work by following it back to the sources, a form of just-in-time verification. In more large-scale data or statistical analysis, queries can be formulated to only apply to records with a certain level of verification, or alternatively the query can highlight lists of pertinent unverified records that the scholar can then verify; this is still laborious, but the expert knows that these entries are precisely those needed to address their research question.

7.8 Summary

In this chapter, we have seen how the growing volume of digital material makes crowdsourcing all but essential if humanities research is to keep pace with the burgeoning source material. However, we have also seen that there is a culture clash between the goal of an authoritative reliable corpus and the perceived potential for inaccuracy, inconsistency, and unreliability of the amateur.

The easiest approach to dealing with this is to confine crowdsourcing to microtasks that only require day-to-day skills such as visual comparisons. Another approach, more suitable for macrotasks, is to increase the quality and confidence in crowdsourced material, for example, traditional dual keying, sanity check rules, or the multi-stage workflows of Distributed Proofreaders (2018).

In *In Concert*, we adopted elements of both of these, albeit for automatic processing and trusted non-experts rather than fully crowdsourced material. However, these were set within a human and digital structure that helped the humanities scholars to retain control of the process. This signposts ways in which crowdsourced material from both microtasks and macrotasks could be similarly included in digital archives so long as their presence is adequately recorded. By making the editorial provenance of data clear, academics can then use their own scholarly judgment as to the reliability of different classes of material and editors for different purposes.

Most crucially, any systems, whether automatic or crowdsourced, need to respect established underlying scholarly values. By so doing we can radically reimagine the processes that lead to the creation of digital archives, but do so in ways that preserve their fundamental academic integrity.

References

- Ackerman, P. (1922). *Catalogue of the Retrospective Loan Exhibition of European Tapestries*, Taylor and Tayloy, NY. <http://www.gutenberg.org/ebooks/57518>.
- Ahmed, E., Ipeirotis, P., & Verykios, V. (2007). Duplicate record detection: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 19(1), 1–16. <https://doi.org/10.1109/TKDE.2007.9>.
- Bashford, C., Cowgill, R., & McVeigh, S. (2000). The Concert Life in Nineteenth-Century London Database, in *Nineteenth-Century British Music Studies*, 2, ed. by J. Dibble and B. Zon (Aldershot: Ashgate, 2000) (pp. 1–12).
- Bell, D. (2004). *Infinite archives, substance* (Vol. 33, No. 3, Issue 105, pp. 148–161). University of Wisconsin Press. <http://www.jstor.org/stable/3685549>.
- Berners, T.L. (1989). Information management: A Proposal. CERN internal report, March 1989, May 1990. <http://info.cern.ch/Proposal.html>.
- Bhattacharya, I., & Getoor, L. (2007). Collective entity resolution in relational data. *ACM Transactions on Knowledge Discovery from Data*, 1(1), 5.
- Bodleian Library (2012/2019). *What's the Score at the Bodleian? Bodleian Library*. Retrieved May 1, 2019, from <http://scores.bodleian.ox.ac.uk>.
- Borges, J. (1946). Del rigor en la ciencia. (tr. 'On Exactitude in Science') *Los Anales de Buenos Aires* 1.3 (March 1946):53.
- Brown, J., & Stratton, S. (1897). *British Musical Biography: a dictionary of musical artists, authors and composers, born in Britain and its colonies*. S.S. Stratton, Birmingham. OCR text: <https://archive.org/details/britishmusicalb00browseable> and data version: <http://www.datatodata.com/in-concert/BMB/>.
- Cheng, J., Teevan, J., Iqbal, S. T., & Bernstein, M. S. (2015). Break it down: A comparison of macro- and microtasks. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, USA (pp. 4061–4064). <https://doi.org/10.1145/2702123.2702146>.
- Concert Life in 19th-Century London database project*, funded by the University of Huddersfield and Oxford Brookes University (1997–2001), and the Arts and Humanities Research Board (UK) and University of Leeds (2001–04).
- Concert Programmes online database. Created 2004–2007. Retrieved September 29, 2018, from <http://www.concertprogrammes.org.uk/about/>.
- Cowgill, R., & Poriss, H. (eds) (2012). *The arts of the prima donna in the long nineteenth century*. Oxford University Press.
- Di Gioia, M., Scannapieco, M. & Beneventano, D. (2010). Object identification across multiple sources. In *Proceedings of the Eighteenth Italian Symposium on Advanced Database Systems, SEBD 2010*, Rimini, Italy, June 20–23, 2010.
- Distributed Proofreaders (2018). *Distributed proofreaders: Preserving history one page at a time*. Retrieved September 02, 2018, from <https://www.pgdp.net/>.
- Dix, A. (2019). *Creativity – understanding and enhancing technical creativity and innovation*. Retrieved November 11, 2019, from <https://alandix.com/creativity/>.
- Dix, A., Beale, R., & Wood, A. (2000). Architectures to make simple visualisations using simple systems. In *Proceedings of the Working Conference on Advanced Visual Interfaces* (pp. 51–60). ACM
- Dix, A., Cowgill, R., Bashford, C., McVeigh, S., & Ridgewell, R. (2014). Authority and judgement in the digital archive. In *Proceedings of the 1st International Workshop on Digital Libraries for Musicology (DLfM '14)*. ACM, New York, NY, USA (pp. 1–8). <https://doi.org/10.1145/2660168.2660171>.
- Dix, A., Cowgill, R., Bashford, C., McVeigh, S., & Ridgewell, R. (2016). Spreadsheets as user interfaces. In *Proceedings of AVI2016*, ACM (pp. 192–195). <https://doi.org/10.1145/2909132.2909271>.

- Dunn, H. (1946). Record linkage. *American Journal of Public Health*, 36(12), 1412–1416. <https://doi.org/10.2105/AJPH.36.12.1412>.
- Fink, F., Schulz, K. U., & Springmann, U. (2017). Profiling of OCR'ed historical texts revisited. In *Proceedings of the 2nd International Conference on Digital Access to Textual Cultural Heritage (DATECH2017)*. ACM, New York, NY, USA (pp. 61–66). <https://doi.org/10.1145/3078081.3078096>.
- Gove, M. (2016). *Sky News interview with Faisal Islam*, 6 June 2016.
- Grove, G. (Ed.). (1900). *A Dictionary of Music and Musicians AD 1450-1880* (Vol. 3). Macmillan.
- Haas, D., Ansel, J., Gu, L., & Marcus, A. (2015). Argonaut: macrotask crowdsourcing for complex data processing. *Proceedings of the VLDB Endowment*, 8(12), 1642–1653. <http://dx.doi.org/10.14778/2824032.2824062>.
- In Concert (2014–2016). Retrieved January 03, 2016 from <http://inconcert.datatodata.com>.
- Leverhulme Trust (2018). *Research Project Grants*. Retrieved September 04, 2018, from <https://www.leverhulme.ac.uk/funding/grant-schemes/research-project-grants>.
- McVeigh, S. (1992–2014). Calendar of London Concerts 1750–1800. (Dataset) Goldsmiths, University of London. <http://research.gold.ac.uk/10342/>.
- Nikolov, A., d'Aquin, M., and Motta, E. (2012). Unsupervised learning of link discovery configuration. In *Proceedings of ESWC'12, Springer, Berlin, Heidelberg* (pp. 119–133). https://doi.org/10.1007/978-3-642-30284-8_15.
- Nurmikko-Fuller, T., Dix, A., Weigl, D. M., & Page, K. R. (2016). In collaboration with in concert: reflecting a digital library as linked data for performance ephemera. In *Proceedings of the 3rd International workshop on Digital Libraries for Musicology (DLfM 2016)*. ACM, New York, NY, USA (pp. 17–24). <https://doi.org/10.1145/2970044.2970049>.
- OpenRefine: Reconciliation Service API. Retrieved September 24, 2018, from <https://github.com/OpenRefine/OpenRefine/wiki/Reconciliation-Service-API>.
- Part 2D: Main Panel D criteria, Panel criteria and working methods, REF2014, Research Excellence Framework. January 2012. <http://www.ref.ac.uk/pubs/2012-01/>.
- Rendle, S. & Schmidt-Thieme. L. (2006). Object identification with constraints. *Data Mining*, 2006 1026–1031. http://www.ismll.uni-hildesheim.de/pub/pdfs/Rendle_SchmidtThieme2006-Object_Identification_with_Constraints.pdf.
- Rusbridge, C. (2007). Arts and Humanities Data Service decision. DCC News, 6 June, 2007. Digital Curation Centre. <http://www.dcc.ac.uk/news/arts-and-humanities-data-service-decision>.
- Scannapieco, M., Tosco, L., Valentino, L., Mancini, L., Cibella, N., Tuoto T., & Fortini, M. (2015). Relais User's Guide – Version 3.0. Technical Report, Italian National Institute of Statistics (Istat). July 2015. <https://doi.org/10.13140/rg.2.1.1332.5922>.
- Schmitz, H., & Lykourentzou, I. (2018). Online sequencing of Non-decomposable macrotasks in expert crowdsourcing. *ACM Transactions on Social Computing*, 1(1), 1. Article 1 (January 2018), 33 p. <https://doi.org/10.1145/3140459>.
- Transforming Musicology*. Retrieved January 03, 2016, from <http://www.transforming-musicology.org>.
- Vobl, T., Gotscharek, A., Reffle, U., Ringlstetter, C., & Schulz, K. U. (2014, May). PoCoTo—an open source system for efficient interactive postcorrection of OCR'ed historical texts. In *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage* (pp. 57–61). ACM. <http://doi.org/10.1145/2595188.2595197>.
- von Ahn, Luis, Maurer, Benjamin, McMillen, Colin, Abraham, David, & Blum, Manuel. (2008). reCAPTCHA: Human-based character recognition via web security measures. *Science*, 321(5895), 1465–1468.
- Wikipedia. (2019). *Arts and humanities data service*. Retrieved January 01, 2019, from https://en.wikipedia.org/wiki/Arts_and_Humanities_Data_Service.

Part III
Macrotasking for Social Good

Chapter 8

“You Can Do It!”—Crowdsourcing Motivational Speech and Text Messages



Roelof A. J. de Vries, Khiet P. Truong, Jaebok Kim and Vanessa Evers

Abstract Recent approaches for technology, that assist or encourage people to change their exercise behavior, focus on tailoring the content of motivational messages to the user. In designing these messages, the mode and style of presentation, e.g., spoken or written and tone of voice, are also thought to play an important role in the effectiveness of the message. We are interested in studying the effects of the content, mode, and style of motivational messages in the context of exercise behavior change. However, we are not aware of any accessible database on motivational messages. Moreover, collecting a large database of spoken and written messages is not a trivial task. Crowdsourcing can be an effective way to collect a large amount of data for all sorts of tasks. Traditionally, crowdsourcing tasks are relatively easy for participants (microtasks). In this work, we use crowdsourcing to collect a large amount of data for more complex tasks (macrotasks): designing motivational messages in text and recording spoken motivational messages. We present and discuss the approach, database and challenges we ran into, and report findings on unsupervised explorations of the emotional expressiveness and sound quality (signal-to-noise ratio, SNR) of the crowdsourced motivational speech.

8.1 Introduction

Recently, there is a growing interest to investigate and develop motivational technology that assists or encourages people to change their behavior (Hekler et al. 2013). This technology can be used to encourage the user, for example, to exercise more by pushing motivational messages to the user on mobile phones (Klasanja and Pratt 2012). Many studies describing the development of their technology do not explain in detail *how* they designed the motivational messages used (Latimer et al. 2010). The framing, content, and designer of motivational messages is an important and not

R. A. J. de Vries (✉)

Biomedical Signals and Systems, University of Twente, Enschede, The Netherlands
e-mail: rajdevries@gmail.com

K. P. Truong · J. Kim · V. Evers

Human Media Interaction, University of Twente, Enschede, The Netherlands

© Springer Nature Switzerland AG 2019

V.-J. Khan et al. (eds.), *Macrotask Crowdsourcing*,

Human–Computer Interaction Series, https://doi.org/10.1007/978-3-030-12334-5_8

a trivial aspect that should be considered carefully when developing motivational or behavior change technology (de Vries et al. 2017a; de Vries 2018). Personalization (e.g., tailoring to the user's personality, Arteaga et al. 2010; de Vries et al. 2016a, 2017b) could for example be a framing method with a positive influence on exercise adherence, but this needs to be investigated in more detail. Furthermore, the mode of presentation and style, e.g., spoken or written and tone of voice, could play an important role in exercise adherence.

In order to study the design and effectiveness of different types of motivational messages for motivational technology, a large database with varying motivational messages in different modes of presentation (i.e., spoken and written) was developed by the authors. For our purpose, namely developing a smartphone application to support exercise behavior change, we decided to gather motivational messages not only in text form but also in spoken form. Rather than generating a small set of messages or relying on experts, we opted for generating a large set of motivational messages by non-experts (peers) through *crowdsourcing*. In our crowdsourcing survey, participants were asked to come up with motivational messages (submitted in *written* and *spoken* form) for a hypothetical person in a given scenario about exercising. This setup allows us to collect a large number of written and spoken motivational messages to study the effectiveness of the message's modality (written vs. spoken), content (themes and topics of the messages relating to the scenarios), and vocal expressivity. In this chapter, we focus on the vocal expressivity of the spoken messages.

Crowdsourcing is usually used for small and easy tasks called microtasks (Cheng et al. 2015). Crowdsourcing written transcriptions, translations or annotations (e.g., Marge et al. 2010; Zaidan and Callison-Burch 2011; Hsueh et al. 2009) is a relatively frequent natural language processing task. However, eliciting spoken data through crowdsourcing seems to be less common and to our knowledge, this is the first effort in using crowdsourcing for the complex task of collecting text-based motivational messages as well as spoken motivational messages. A challenge for a complex task, also called a (non-decomposable) macrotask (Schmitz and Lykourantzou 2018), like this is evaluating the quality of the workers' output, because no ground truth is available (Haas et al. 2015). Crowdsourcing spoken messages brings along additional challenges: loss of control over the recorded sound quality and the speaking style of the participant are among those challenges. Participants will have different types of microphones with varying qualities and there is no knowing to what extent the spoken material actually reflects a motivational speaking style after listening to all the audio recorded. Despite these challenges, it would be useful to explore the feasibility of acquiring spoken data through crowdsourcing involving variations in speaking styles (i.e., motivational) that enables paralinguistic research, which is still a rather uncovered area in crowdsourcing.

In this chapter, we present our approach to crowdsourcing spoken (and written) motivational messages and present our collected corpus. We discuss how we designed the data collection and we report on (1) the audio quality (SNR) of the crowdsourced audio material and (2) an initial, unsupervised exploration of the acoustical feature space of motivational speech. We describe related work in Sect. 8.2 and present our

data collection effort in Sect. 8.3. We report on an preliminary exploration of the quality and acoustics of motivational speech in Sect. 8.4 and discuss the conclusions and future research in Sect. 8.5.

8.2 Related Work

We explain relevant psychological concepts used in our study and discuss previous related work.

8.2.1 *Motivation and Exercise Behavior Change*

According to the Transtheoretical Model (TTM, Prochaska and DiClemente 1983), people, who change their exercise behavior for example, will go through certain *stages of change*. These five stages of change classify people into progressing stages of behavior change as follows: Precontemplation (not considering change), Contemplation (thinking of change), Preparation (preparing for change), Action (actively making changes), and Maintenance (maintaining the change). We expect that motivational messages attuned to the stages of change a user is in will be more effective for exercise adherence. However, in an evaluation of the text version of the spoken motivational messages described in this chapter, we found that the way people *rate* the messages on how motivating they are does not always match the expectation of what messages should be most effective for the stage of change they are in (more details are reported in de Vries et al. (2016b)).

8.2.2 *Crowdsourcing Text and Speech*

Over the last few years, researchers have been using crowdsourcing platforms such as Amazon Mechanical Turk (AMT) for various natural language processing (NLP) tasks. Callison-Burch and Dredze (2010) and Parent and Eskenazi (2011) summarize the kind of NLP tasks commonly addressed which include, among others, transcriptions of spoken language (Marge et al. 2010), producing and evaluating (machine) translations (Zaidan and Callison-Burch 2011; Callison-Burch 2009), and sentiment labeling (Hsueh et al. 2009). These tasks usually involve assessing text or spoken data. Crowdsourcing platforms can also be used to *acquire* spoken language data. Although challenging (for example, there is no way to control the microphone type, distance or noise level), collecting spoken language data through crowdsourcing can be a cost- and time-effective way to gather large amounts of speech data under realistic conditions. Recently, efforts to collect speech data through crowdsourcing have been undertaken involving tasks such as reading aloud street addresses (McGraw

et al. 2010), having conversations with a spoken dialogue system (McGraw et al. 2010), narrating Wikipedia articles for use by blind or illiterate users (Novotney and Callison-Burch 2010), reading aloud sentences in under-resourced languages (Lane et al. 2010), and annotating photos through spoken descriptions for a voice search system (McGraw et al. 2011). Challenges discussed in these studies include loss of (quality) control and also technical challenges since incorporating a web-based audio collection framework in crowdsourcing platforms such as AMT is not straightforward. Studies on the prosody of motivational speech, with the exception of a recent study by Skutella et al. (2014) are rare. In instructor–trainee indoor cycling sessions they found, among other things, a high frequency of prominent, accented words fulfilling a coordinative and informative function. We are aware of only one related study on collecting motivational messages, by Coley et al. (2013), where written *text* messages were crowdsourced to encourage people to quit smoking. With our effort of crowdsourcing motivational speech and text messages, we aim to address this lack of data and research and demonstrate the feasibility of crowdsourcing spoken motivational messages.

8.2.3 *Defining Macrotasks and Microtasks*

Macrotasking, as defined by this book, refers to complex and often creative crowd work, which may or may not be decomposable to microtask level, but which differs from microtasking in that it requires more worker time, can accept free-form worker input (i.e., not only multiple-choice standardized input), and its quality needs to be, at least partially, determined through subjective evaluation, for example peer review. Microtasks, in contrast, are small tasks that are easily performed. Microtasks are frequently used in crowdsourcing (Cheng et al. 2015).

Considering the tasks mentioned in the related works discussed in the previous section in light of this definition of macro and microtasks, all of those tasks mentioned could be considered microtasks, although for some this is only because they are decomposed to microtask level. Narrating articles (Novotney and Callison-Burch 2010), reading aloud sentences (Lane et al. 2010), or transcribing spoken language (Marge et al. 2010) is relatively easy and straightforward and therefore fits the microtasks definition well. However, producing and evaluating (machine) translations (Callison-Burch 2009; Zaidan and Callison-Burch 2011) and sentiment labeling (Hsueh et al. 2009), depending on the difficulty of the text, can require some cognitive effort. Moreover, having conversations with a spoken dialogue system (McGraw et al. 2010) and annotating photos through spoken descriptions for a voice search system (McGraw et al. 2011) can also require quite some cognitive effort depending on the dialogue or the photo. For these tasks, it seems that what qualifies them for microtasks is that these tasks were decomposed to the simplest level, such as describing only one photo, or have one short dialogue, and in that way they require very little worker time. On the other hand, these tasks could also qualify for macrotasks because they require free-form input, the quality of the input needs to be determined through subjective evaluation, and the tasks are not easily performed.

Applying this definition to the task we designed, our crowdsourcing task can be considered a macrotask. Our crowdsourcing task required creativity and quite some worker time (participants were asked to come up with multiple motivational messages for a hypothetical person in a given scenario about exercising, see Tables 8.1 and 8.2), accepted only free-form worker input (the participants had to design all the messages from scratch), and the quality was partially determined through subjective evaluation (more details on our evaluation are reported in de Vries et al. 2016b). This is also what makes a macrotask like this challenging, because there is no ground truth available for evaluating the quality of the workers’ output (Haas et al. 2015). Moreover, it is challenging because we decided to gather motivational messages not only in text form but also in spoken form. Also, crowdsourcing spoken messages brings along additional challenges (e.g., loss of control over the recorded sound quality and speaking style of the participant).

On the other hand, our macrotasks could be decomposed into smaller tasks by asking participants for only one motivational message each. In this way, the task would require less participant time and could arguably move toward a microtask. However, this task would then be non-decomposable and still require a certain creativity of the participant (to come up with a motivational message) and the quality of the message would still be determined through subjective evaluation. Moreover, for the purpose of our data collection, we were also interested in variation in the motivational messages, which is stimulated by asking participants for multiple motivational messages, in that sense the task was non-decomposable. Another facet to consider is the complexity of the tasks. According to Schmitz and Lykourantzou (2018, p. A:7): “Macrotasks are almost always complex, in that they require multiple interconnected knowledge domains ...”. Our task however, is not so complex or difficult that it requires worker training, in fact, the task is purposefully crowdsourced

Table 8.1 One of the five stage of change scenarios (Contemplation) and the macrotask of designing multiple motivational messages for specific time frames (with one collected example)

One of the stage of change scenarios: Contemplation

Contemplation: “Consider a middle-aged person, with a steady personal life and solid friend foundation. This person lacks regular exercise in his/her daily life, but has been thinking about starting to exercise regularly and wonders if he/she will be able to do it. This person is opting to start in the next 6 months”

Long: “Imagine you have to provide this person with motivational messages during a long period of time (for example, 1 year) and these messages take into account the current exercise habits as described. These messages would be provided every other week (for example, week 1 and week 3 of every month). What would be 3 messages you can think of?” **Example:** “You have to start somewhere”

Short: “Imagine you have to provide this person with motivational messages during a short period of time (for example, 1 month) and these messages take into account the current exercise habits as described. These messages would be provided three times a week (for example Monday, Wednesday and Friday). What would be 3 messages you can think of?”

Table 8.2 One of the three running performance scenarios (Running too fast) and the macrotask of designing multiple motivational messages for specific points of time in a run (with one collected example)

One of the running performance scenarios: Running too fast

Too fast: “Consider a person during an actual exercise, for example running, he/she is well under way in his/her run but for the purpose of a good exercise it would be best if he/she decreases the intensity of the run”

During: “Imagine you have to provide this person with motivational messages during this session of physical activity and these messages would be provided to encourage and motivate this person to decrease the intensity during the run. What would be 3 motivating messages you can think of?”

After: “Consider the same person after the exercise (the run), he/she has exercised and so he/she is done, but did not succeed in decreasing the intensity of the run, despite the motivational messages, and is now cooling down. Although disappointing at this moment, running regularly is what is most important. What would be 3 motivating messages you can think of?” **Example:** “Great run, but watch your speed next time”

Before: “Consider the same person before his/her next exercise (the run). In the last run it would have been better to have had a lower intensity. This person decides to run again, partially because of the messages during his/her cooling down the other day, and is ready to start. What would be 3 motivating messages you can think of?”

to reach people who do not have the domain knowledge to design expertise driven motivational messages (designing motivational messages is the task), but who design motivational messages from their (potentially limited) own experience. In that sense, our task does not fit the general complexity criterion of macrotasks.

8.3 Data Collection

We describe how we designed our macrotasks and collected a database of spoken and written motivational messages through crowdsourcing.

8.3.1 Participants

We recruited participants via AMT. The requirements were that they needed to have completed more than a 1000 tasks on AMT, more than 98% of their tasks needed to be approved successfully, and they needed to be located in the US. These requirements ensured that we would have participants who were experienced and serious in filling in questionnaires, and that they had good proficiency in English (95% of the recruited participants reported “very good” for their self-assessed proficiency of English). The sample size consists of 500 people. Of these, 17 were excluded because their data is incomplete or have numerous outliers. Then, another 19 were excluded because they

have missing audio files (recording audio was encouraged but not strictly required to finish the survey). The final sample for spoken messages includes 464 participants (246 male). All but 4 participants were native English speakers. The minimum age was 18 and the maximum was 68. The average age was 30.93 ($SD = 9.13$) and the median 29.0.

8.3.2 *Method*

The macrotask for the participants throughout this survey was to come up with motivational messages to motivate certain people in different scenarios (in a randomized order). Since one of the features of our intended smartphone application is the use of motivational messages tailored to the stage of change, scenarios were manipulated based on the stages of change. See Table 8.1 for examples that describe a person in a situation corresponding to a certain stage of change. Participants were asked to come up with 6 different messages to motivate this person to exercise more, 3 for the **short** and 3 for the **long** term.

Another intended feature of our smartphone application is to provide motivational feedback about the quality of exercise. Hence, the second manipulation involved the running performance (running too fast, too slow or exactly right) of the person described, see Table 8.2 for example. Participants were asked to come up with 9 different motivational messages: 3 for **before**, 3 for **during**, and 3 for **after** a running session.

8.3.3 *Implementation*

Although we used AMT to enlist participants, the survey itself could not be embedded in AMT due to technical constraints with collecting audio. We needed to prompt the participants in the survey with the written motivational messages they had come up with earlier (and not predefined prompts) to record them on our web application outside the survey. Because we only found an option with static (predefined) prompts in AMT, we had to come up with a workaround. We used a relatively easy workaround with SurveyMonkey¹ where there is a possibility to use answer text boxes as future variables (prompts). In the web survey, this allowed us to refer to the future variable name identifier (i.e., a number in front of the to-be-instantiated variable). For the crowdsourced speech data acquisition, we set up a web application called the WAMI recorder² with a Google App Engine as described in McGraw (2013)³ and from SurveyMonkey we referred the participants to this page to record their motivational

¹<https://surveymonkey.com>.

²<https://wami-recorder.googlecode.com>.

³<https://wami-gapp.googlecode.com>.

messages. All audio files (~7000) were stored in the Google App Engine in separate folders for each participant and were automatically retrieved via a script. However, the link between participant id and audio id was lost which meant that we needed to manually link each participant to the correct folder through their matching written motivational messages.

8.3.4 Measures

In addition to basic demographic information, participants were asked to fill in a 1-item stage of change measure for exercise (Norman et al. 1998), the Godin Leisure-Time Exercise Questionnaire (Godin and Shephard 1997), a 30-item processes of change measure for exercise (Nigg et al. 1999), an 18-item self-efficacy measure for exercise (Benisovich et al. 1998), a 10-item decisional balance measure for exercise (Nigg et al. 1998),⁴ and the 50-item IPIP personality questionnaire.⁵ These measures are not reported on in this work.

8.3.5 Procedure

Participants signed up on AMT where they were informed of their compensation, goal of the survey and estimated time cost. They were also asked to check whether their browser and microphone worked in a test version of the WAMI recorder. Participants could then decide to proceed to the survey on SurveyMonkey where the consent form was presented. Next, participants were asked to fill in demographics and then the data collection started where they were presented with various scenarios and were asked to come up with motivational messages in written form. Subsequently, participants were asked to vocally express and record the motivational messages (on a different webpage) that they had just written. They were shown the text they had just entered and were asked to repeat the message orally as they intended it. Finally, participants were asked to fill in the questionnaires as described in Sect. 8.3.4. After completion, participants were debriefed about the detailed goals of this survey and given a completion code to fill in on AMT to receive payment. On average, the survey took about 45 min to complete. Participants were paid 3 US dollars for their participation (Table 8.3).

⁴All TTM measures adopted from <http://www.uri.edu/research/cprc/measures.htm>.

⁵Adopted from <http://ipip.ori.org/>.

Table 8.3 Descriptive statistics of the messages collected

N = 6909	Mean	Std	Median	Min	Max
Duration (s)	4.5	2.0	4.2	0.9	32.7
Number of words	9.0	5.6	8	1	97

8.4 Results

One of the main goals of this study was to collect motivational speech, but also to explore its acoustical characteristics in terms of sound quality and emotional expressiveness. We collected a total of 6960 (464×15) motivational messages. Using simple voice activity detection, we discarded 51 messages which did not seem to contain voice at all. First, we explore the sound quality through an analysis of Signal-to-noise ratio (SNR). Second, we analyze how feature vectors of the motivational speech are distributed in a feature vector space of emotional speech: what kind of emotion does motivational speech resemble acoustically? To the best of our knowledge, there is no other motivational speech corpus that we can use as a reference in order to validate our findings. Additionally, we do not assume whether the motivational speech collection contains spontaneous or acted emotional speech data. Therefore, we use both spontaneous and acted available emotional speech corpora as training data for our analyses. We selected the SEMAINE corpus (natural emotional speech) (McKeown et al. 2012) and the LDC Emotional Prosody Speech corpus (acted emotional speech) because of their relatively large size and variety of emotional categories.

8.4.1 SNR of the Motivational Speech Corpus

We estimate the SNR of each spoken message following a method by Hirsch (1993) using voice activity detection (VAD) and assume that each speech sample already contains some noise. Figure 8.1 illustrates the distribution of SNR for three different corpora. Our motivational speech corpus shows a median of 16.97 and mean of 16.69 ± 11.68 , which are considered not optimal for automatic speech recognition (ASR) (Gong 1995; Benzeghiba et al. 2007). The SNR of the motivational speech corpus is lower than that of the other corpora considered (Kruskal–Wallis test: $\chi_2(2)$, $p < 0.0001$, followed by Nemenyi pairwise comparison $p < 0.0001$).

8.4.2 Emotional Feature Vector Space Using LDC and SEMAINE Corpus

Since we used crowdsourcing to collect a large amount of motivational speech data, we could not control for the speaking style of the participants. Moreover, there have been no studies yet (to the best of our knowledge) into prosodic characteristics of

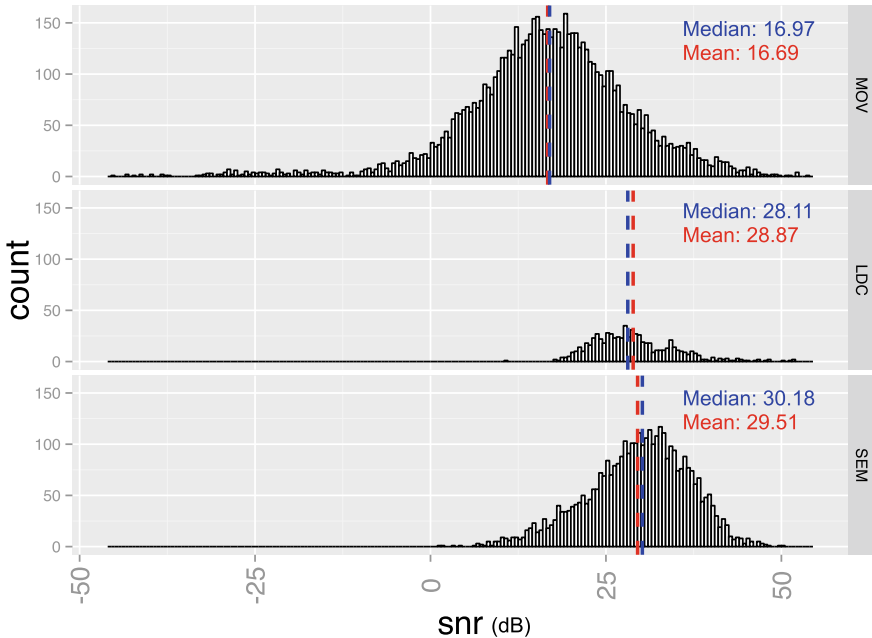


Fig. 8.1 Histogram of SNR of the speech corpora (MOV: the motivational speech corpus, LDC: the LDC Emotional Prosody Speech corpus, and SEM: the SEMAINE corpus)

motivational speech. This makes evaluation difficult. Although we can speculate that motivating people can be done by signaling positive and aroused emotions (Skutella et al. 2014), this is not verified yet. Hence, because of the relatively large amount of speech data and lack of knowledge into prosodic characteristics of motivational speech, we carried out an unsupervised cluster analysis that is exploratory of nature.

Clustering We built clusters (K-means) of each available emotional category in the feature space and investigated how close the feature vectors are to the centers of the clusters. We selected 5 representative emotional categories available in both corpora selected: neutral, happiness, anger, sadness, and boredom (Kwon et al. 2003; Huang and Ma 2006). For the SEMAINE corpus (McKeown et al. 2012), we extracted only speech segments from the users interacting with a human operator (who is playing an emotional character) that is thought to be more spontaneous. The SEMAINE corpus provides only continuous affective ratings, not discrete emotional categories. In order to map these continuous valence and arousal ratings to discrete emotional categories, we used the landmarks of the valence and the arousal dimensions as provided in FEELTRACE (Cowie et al. 2000). We calculated the Euclidean Distance between the landmarks and the values of the valence and arousal dimensions of each segment and assigned the emotional category with the smallest distance to the valence and arousal values. Lastly, we extracted segments by using VAD and time-alignment labels. Table 8.4 summarizes the emotional speech data used to build the emotional feature space.

Table 8.4 Data used to build an emotional feature vector space (No.: number of segments, F: female, M: male, A: arousal, V: valence)

Categories	No. LDC		No. SEMAINE		Landmarks	
	F	M	F	M	A	V
Neutral	34	46	1380	1314	0.00	0.00
Happiness	111	69	253	310	0.74	0.52
Anger	78	61	111	224	-0.77	0.75
Sadness	97	64	32	9	-0.7	-0.48
Boredom	88	90	219	290	-0.43	-0.48

Table 8.5 Normalized mean (standard deviations) of distances between motivational speech and emotional models

Categories	Neutral	Anger	Sadness	Happiness	Boredom
Female	0.24 (0.11)	0.25 (0.12)	0.24 (0.11)	0.24 (0.12)	0.19 (0.11)
Male	0.30 (0.11)	0.33 (0.11)	0.28 (0.12)	0.33 (0.11)	0.25 (0.10)

Feature space To build the emotional feature vector space, we extracted low-level features including energy (RMS), 12 Mel-Frequency Cepstrum Coefficients (MFCCs), prosody (F0, voice probability, zero-crossing rate), and voice quality related features (jitter, shimmer, harmonics-to-noise ratio) from only the voiced parts obtained with VAD. Feature vectors were extracted within frames of 20 ms with a Hamming window by using `openSMILE` (Eyben et al. 2010). We used only mean values of each features to construct clusters in the feature space. Since we do not know which features are dominantly related to motivational speech, we normalized all feature values by the use of the maximum and minimum values on the feature to scale them in a range of [0.0, 1.0] (de Souto et al. 2008). We found a center for each emotional category by calculating the minimum of total Euclidean distances between the center and other vectors. We normalized the distances between the motivational speech vectors and the centers of the emotional models in the same way we did for the features.

Acoustic similarity Table 8.5 presents the means of normalized distances between motivational speech feature vectors and the centers of emotional categories. For both female and male models, we can observe that the motivational speech feature vectors seem to show more acoustic similarity with boredom models than with any other models (Kruskal–Wallis test: $\chi_2(4)$, $p < 0.0001$, followed by Nemenyi pairwise comparison $p < 0.0001$). Especially, in male models, all categories show differences with significance of $p < 0.0001$ between each other except for the pair of happiness and anger.

8.5 Discussion and Conclusion

In this chapter, we presented our text and speech dataset of motivational messages collected through a crowdsourcing macrotask survey. With this data collection effort, we aimed to address the gap in both motivational technology, where datasets of motivational messages are mostly expert-written, not personalized, and relatively small, as well as in speech science, where corpora of motivational speech do not exist yet. Macrotasks, as defined by this book, refers to complex and often creative crowd work, requires more worker time, can accept free-form worker input, and its quality needs to be, at least partially, determined through subjective evaluation. Evaluating macrotasks is a challenge, because there is no ground truth available to evaluate the quality of the workers' output. We used crowdsourcing for a relatively new type of macrotask: eliciting motivational text and speech messages. This task required creative work, a long amount of worker time, free-form input, and subjective evaluation. However, the task was not necessarily very complex in that it required a lot of knowledge domains. A first unsupervised exploration of the acoustic feature space of the acquired motivational speech data was carried out which showed acoustic similarity to mostly low aroused and neutral emotional feature spaces. An SNR analysis showed relatively low SNR values by ASR standards, but we still believe that a large amount of our speech data can be used for paralinguistic research. Our study serves as a good example of how macrotasks in crowdsourcing can be used to for creative elicitation tasks, such as collecting a difficult but context-relevant text and speech dataset of crowd-designed motivational messages for cross-disciplinary use.

Although crowdsourcing seems to be a relatively easy and quick way to acquire a large amount of text and speech data, there are some limitations that one should take into account when using crowdsourcing macrotasks, in particular for a complex macrotask like speech data acquisition, see also McGraw et al. (2010), Parent and Eskenazi (2011) who discuss these limitations as well. From a practical point of view, acquiring speech through well-known crowdsourcing platforms is rather cumbersome and requires some workarounds: browser restrictions, the need to prompt the participants to read aloud what they had previously entered in text, and access to the audio files recorded lead to some cumbersome workarounds which deserve some more elegant solutions in the future. Content-wise for this specific macrotask, the unpredictability of the quality of the acquired audio is still a challenge, both the sound quality and the quality of the desired task, i.e., generating (high-quality) motivational speech. Although a comparison to existing acoustic models might give one a first insight into what the acquired speech might sound like, subsequent analyses such as perceptual rating studies are still needed for confirmation. This need for further evaluation is a general problem for macrotasks.

For future research, we will evaluate the effectiveness of the motivational text and speech messages collected through several user studies. We intend to analyze the messages for linguistic and acoustical patterns in relation to effectiveness and personalized variables such as personality and stages of change. Furthermore, the

dataset might be of interest to researchers working on speech synthesis and natural language generation: imagine an application that automatically generates motivational text and speech messages tailored to the user. Despite some limitations, we believe that our data collection effort also creates many cross-disciplinary and fruitful research opportunities.

References

- Arteaga, S. M., Kudeki, M., Woodworth, A., & Kurniawan, S. (2010). Mobile system to motivate teenagers' physical activity. In *Proceedings of the 9th International Conference on Interaction Design and Children* (pp. 1–10). ACM.
- Benisovich, S., Rossi, J., Norman, G., & Nigg, C. (1998). Development of a multidimensional measure of exercise self-efficacy. *Annals of Behavioral Medicine*, 20(suppl).
- Benzeghiba, M., De Mori, R., Deroo, O., Dupont, S., Erbes, T., Jouvet, D., et al. (2007). Automatic speech recognition and speech variability: A review. *Speech Communication*, 49(10), 763–786.
- Callison-Burch, C. (2009). Fast, cheap, and creative: Evaluating translation quality using Amazon's Mechanical Turk. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1, ser. EMNLP '09* (pp. 286–295).
- Callison-Burch, C., & Dredze, M. (2010). Creating speech and language data with Amazon's Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk* (pp. 1–12).
- Cheng, J., Teevan, J., Iqbal, S. T., & Bernstein, M. S. (2015). Break it down: A comparison of macro-and microtasks. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (pp. 4061–4064). ACM.
- Coley, H. L., Sadasivam, R. S., Williams, J. H., Volkman, J. E., Schoenberger, Y.-M., Kohler, C. L., Sobko, H., Ray, M. N., Allison, J. J., Ford, D. E., Gilbert, G. H., & Houston, T. K. (2013). Crowdsourced peer- versus expert-written smoking-cessation messages. *American Journal of Preventive Medicine*, 45(5), 543–550. <http://www.ncbi.nlm.nih.gov/pubmed/24139766>.
- Cowie, R., Douglas-Cowie, E., Savvidou*, S., McMahon, E., Sawey, M., & Schröder, M. (2000). 'FEELTRACE': An instrument for recording perceived emotion in real time. In *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion* (pp. 19–24).
- de Souto, M. C. P., de Araujo, D. S., Costa, I. G., Soares, R. G., Ludermir, T. B., & Schliep, A. (2008). Comparative study on normalization procedures for cluster analysis of gene expression datasets. In *IEEE International Joint Conference on Neural Networks. IJCNN 2008 (IEEE World Congress on Computational Intelligence)* (pp. 2792–2798). IEEE.
- de Vries, R. (2018). *Theory-based and tailor-made: Motivational messages for behavior change technology*. PhD dissertation, Human Media Interaction, Netherlands.
- de Vries, R. A. J., Truong, K. P., & Evers, V. (2016a). Crowd-designed motivation: Combining personality and the transtheoretical model. In *Persuasive technology* (pp. 41–52). Berlin: Springer.
- de Vries, R. A. J., Truong, K. P., Kwint, S., Drossaert, C. H. C., & Evers, V. (2016b). Crowd-designed motivation: Motivational messages for exercise adherence based on behavior change theory. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 297–308). ACM.
- de Vries, R. A., Zaga, C., Bayer, F., Drossaert, C. H., Truong, K. P., & Evers, V. (2017a). Experts get me started, peers keep me going: Comparing crowd-versus expert-designed motivational text messages for exercise behavior change. In *11th EAI International Conference on Pervasive Computing Technologies for Healthcare, PervasiveHealth*. ACM.
- de Vries, R. A., Truong, K. P., Zaga, C., Li, J., & Evers, V. (2017b). A word of advice: How to tailor motivational text messages based on behavior change theory to personality and gender. *Personal and Ubiquitous Computing*, 21(4), 675–687.

- Eyben, F., Wöllmer, M., & Schuller, B. (2010). Opensmile: The munich versatile and fast open-source audio feature extractor. In *Proceedings of the International Conference on Multimedia* (pp. 1459–1462). ACM.
- Godin, G., & Shephard, R. (1997). Godin leisure-time exercise questionnaire. *Medicine and Science in Sports and Exercise*, 29(6s), S36.
- Gong, Y. (1995). Speech recognition in noisy environments: A survey. *Speech Communication*, 16(3), 261–291.
- Haas, D., Ansel, J., Gu, L., & Marcus, A. (2015). Argonaut: Macrotask crowdsourcing for complex data processing. *Proceedings of the VLDB Endowment*, 8(12), 1642–1653.
- Hekler, E. B., Klasnja, P., Froehlich, J. E., & Buman, M. P. (2013). Mind the theoretical gap: Interpreting, using, and developing behavioral theory in HCI research. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 3307–3316).
- Hirsch, H. G. (1993). *Estimation of noise spectrum and its application to SNR-estimation and speech enhancement*. International Computer Science Institute.
- Hsueh, P.-Y., Melville, P., & Sindhvani, V. (2009). Data quality from crowdsourcing: A study of annotation selection criteria. In *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing, ser. HLT '09* (pp. 27–35).
- Huang, R., & Ma, C. (2006). Toward a speaker-independent real-time affect detection system. In *Proceedings of the International Conference on Pattern Recognition (ICPR), 1*, 1204–1207.
- Klasnja, P., & Pratt, W. (2012). Healthcare in the pocket: Mapping the space of mobile-phone health interventions. *Journal of Biomedical Informatics*, 45(1), 184–198.
- Kwon, O.-W., Chan, K., Hao, J., & Lee, T.-W. (2003). Emotion recognition by speech signals. In *Proceedings of Interspeech* (pp. 125–128).
- Lane, I., Waibel, A., Eck, M., & Rottmann, K. (2010). Tools for collecting speech corpora via Mechanical-Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk* (pp. 184–187).
- Latimer, A. E., Brawley, L. R., & Bassett, R. L. (2010). A systematic review of three approaches for constructing physical activity messages: What messages work and what improvements are needed? *The International Journal of Behavioral Nutrition and Physical Activity*, 7, 36.
- Marge, M., Banerjee, S., & Rudnicky, A. I. (2010). Using the Amazon Mechanical Turk for transcription of spoken language. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5270–5273).
- McGraw, I. (2013). Collecting speech from crowds. In *Crowdsourcing for speech processing: Applications to data collection, transcription and assessment* (pp. 37–71).
- McGraw, I., Glass, J., & Seneff, S. (2011). Growing a spoken language interface on Amazon Mechanical Turk. In *Proceedings of Interspeech* (pp. 3057–3060).
- McGraw, I., Lee, C., Hetherington, L., Seneff, S., & Glass, J. (2010). Collecting voices from the cloud. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)* (pp. 1576–1583).
- McKeown, G., Valstar, M., Cowie, R., Pantic, M., & Schroder, M. (2012). The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Transactions on Affective Computing*, 3(1), 5–17.
- Nigg, C., Rossi, J., Norman, G., & Benisovich, S. (1998). Structure of decisional balance for exercise adoption. *Annals of Behavioral Medicine*, 20, S211.
- Nigg, C., Norman, G., Rossi, J., & Benisovich, S. (1999). Processes of exercise behavior change: Redeveloping the scale. *Annals of Behavioral Medicine*, 21, S79.
- Norman, G., Benisovich, S., Nigg, C., & Rossi, J. (1998). Examining three exercise staging algorithms in two samples. In *19th Annual Meeting of the Society of Behavioral Medicine*.
- Novotney, S., & Callison-Burch, C. (2010). Crowdsourced accessibility: Elicitation of Wikipedia articles. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk* (pp. 41–44).

- Parent, G., & Eskenazi, M. (2011). Speaking to the crowd: Looking at past achievements in using crowdsourcing for speech and predicting future challenges. In *Proceedings of Interspeech* (pp. 3037–3040).
- Prochaska, J. O., & DiClemente, C. C. (1983). Stages and processes of self-change of smoking: Toward an integrative model of change. *Journal of Consulting and Clinical Psychology, 51*(3), 390.
- Schmitz, H., & Lykourantzou, I. (2018). Online sequencing of non-decomposable macrotasks in expert crowdsourcing. *ACM Transactions on Social Computing, 1*(1), 1.
- Skutella, L. V., Sssenbach, L., Pitsch, K., & Wagner, P. (2014). The prosody of motivation: First results. In *Proceedings of ESSV (Elektronische Sprachsignalverarbeitung)*.
- Zaidan, O. F., & Callison-Burch, C. (2011). Crowdsourcing translation: Professional quality from non-professionals. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, ser. HLT '11, 1*, 1220–1229.

Chapter 9

Crowdsourcing Real-World Feedback for Human–Computer Interaction Education



Fernando Loizides, Kathryn Jones, Carina Girvan, Helene de Ribaupierre, Liam Turner, Ceri Bailey and Andy Lloyd

Abstract In this chapter we investigate using real-world feedback to compliment academic feedback during a course on mobile development using HCI methods. Students used crowdsourcing and macro tasking to recruit suitable end-users in order to generate feedback. During the study, we uncovered benefits and disbenefits for both staff and student stakeholders, motivations and blockers that readers and others planning to apply this method should be aware of. We report on practical matters of scalability, legal and governance issues that arise. Overall, the process proposed in the chapter produces a greatly enhanced experience for students and improves the richness of the feedback as well as the authenticity of the end-user testing experience. Challenges that are faced include incorporating this process within an academic environment with matters such as the liability of the university towards externalising student work. Overall, we were also surprised to see that harsh criticism was not taken negatively by students but was a source of motivation to improve.

F. Loizides (✉) · K. Jones · H. de Ribaupierre · L. Turner
School of Computer Science and Informatics, Cardiff University, Cardiff, UK
e-mail: LoizidesF@cardiff.ac.uk

K. Jones
e-mail: jonesk90@cardiff.ac.uk

H. de Ribaupierre
e-mail: deRibaupierreH@cardiff.ac.uk

L. Turner
e-mail: TurnerL9@cardiff.ac.uk

C. Girvan · C. Bailey
School of Social Sciences, Cardiff University, Cardiff, UK
e-mail: GirvanC@cardiff.ac.uk

C. Bailey
e-mail: BaileyCM1@cardiff.ac.uk

A. Lloyd
Centre for Education Support and Innovation, Cardiff University, Cardiff, UK
e-mail: LloydA@cardiff.ac.uk

9.1 Introduction and Motivation

Collecting feedback is valuable in the development process for products such as software, as it enables a degree of evaluation before going to mass market (e.g. Alpha and Beta product releases). By undertaking this process as a macrotask, whereby feedback is amassed as a collective body of information from many ‘specialised’ individuals, this process becomes richer and more useful. The use of the term *macrotask* has been defined in several contexts. We define macrotasking to be a complete task (from start to finish) that requires a certain level of expertise and/or collaboration. We distinguish this from the term *microtasking* which we define as a smaller scale task, which can be performed to contribute to the success of a larger macrotask, and often requires no expertise or collaboration but necessitates human perception.¹ A macrotask could then be broken down into a series of microtasks (Cheng et al. 2015) which require ‘enabling much of the complexity of traditional knowledge work’ to bring them back together to complete the macrotask (Haas et al. 2015).

In higher education, Human–Computer Interaction (HCI) courses follow traditional methods of assessment; namely, the lecturer provides a predefined set of criteria for students, who create applications using the taught interaction design techniques, such as user-centred design. Ultimately, the artefacts produced are assessed within the constraints of a rubric (Jonsson 2014), which rely on the expert discretion of the lecturer. While this is the established norm in HCI education, it is an approach which is at odds with the development processes students will encounter when they enter the world of work. Thus, accurate reflection and training for early stage HCI practitioners is often not provided to the trainee developers and User Experience (UX) students.

The challenge is how to provide students with authentic learning experiences, which mirror and prepare them for the feedback mechanisms they will encounter in the workplace. Within the authors’ institution, like many in the UK and elsewhere, there are calls from within the institution, both at a management level and from students for ‘better’ feedback from lecturers. While on the one-hand, lecturers are keen to provide learners with authentic, real-world feedback, they are constrained by structural limitations imposed by the environment in which traditional teaching and learning occurs, such as timing, volume and turn-around times of assessment and feedback. To address the challenge of providing students with an authentic experience within the structural limitations of the institution, this chapter describes and evaluates a pilot study of an innovative approach to assessment and feedback of an HCI application design task. By sourcing feedback from a community of users that have domain knowledge and are technology literate, we can spread the load of assessment and feedback while also gaining valuable insights into whether the application is useful in the real world. Issues regarding the management of such an ambitious crowdsourcing task are highlighted, such as the privacy implications of releasing prototype software by ‘amateur’ HCI practitioners. Furthermore, strong

¹<https://blog.gems.org/gems-what-are-micro-tasks-and-why-are-they-important-87b5b35eef3b>— Accessed December 2018.

social, psychological and legal impacts of such an approach to crowdsourcing arise and are also presented. Finally, the findings demonstrate the importance of reflection through crowdsourcing, a unique and interesting aspect that is largely overlooked in traditional microtasking, by providing structured opportunities for students to reflect upon and where appropriate respond to crowdsourced feedback.

We begin by situating the reader within the related work, followed by a brief description of our arrangements to expose students to end-user feedback. We then present both students' and lecturers' perspectives in a themed framework which the reader can benefit from while planning his or her own assessment criteria. Finally, following on from one of our findings, we also present a framework for an automated natural language support system, to allow for scaling of such a user feedback task.

9.2 Current Context of Assessment and Feedback in HCI Courses

Assessment and feedback have long been recognised as the area of their learning experience that students are most often least satisfied with. For example, on average, over the last three years, undergraduate student satisfaction with assessment and feedback in computing science is about ten percent lower than overall satisfaction.² In seeking to address and enhance provision in this area, a range of research has been undertaken to identify why lower satisfaction continues to be reported in this area. From this, a number of factors have been identified and a range of strategies and initiatives implemented to enhance the student experience in this area, which are described below. While these have led to some 'marginal improvements' in student satisfaction, it remains an area of particular concern within higher education (Soilemetzidis et al. 2014).

The range of factors identified that can impact on student satisfaction also highlight the linked and integrated nature of assessment and feedback, interventions in one area often impacting on practice in others, which in turn can lead to ongoing student unhappiness (Jessop and Thomas 2016). Thus, there is a need for interventions to be put in place that address all of the areas that can make a difference, such as the design and format of assessments, the ability of students to understand and interpret the feedback they receive, and the opportunities for students to engage in dialogue so that they can utilise feedback to improve learning.

This is especially true within the context of an HCI or User Experience class. Although advocating for user-centred design, there is often minimal or no actual user involvement within the assessment and feedback criteria. Logistically this is understandable, as there are institutional and practical barriers to employing real life users to 'mark' or provide feedback within a lesson. Factors such as intellectual property and lack of users are often an issue. Some lecturers may work around this

²See <https://unistats.ac.uk/> [Accessed December 2018] for a summary of the outcomes arising from the National Student Survey in specific subject areas.

by using different techniques to involve some sort of user feedback. For example, by employing industry professionals to set software needs which are turned into course-work, the National Software Academy at Cardiff University³ is able to collaborate with industrial partners to give feedback to students as to the coverage of business and functional requirements needed. There is also a limited ability to be able to give the product to the end users; namely, when the product that is being developed is to be used internally by the company employees. A more common approach to introducing user-centred design or evaluation for feedback, is that of asking students to find users themselves. This often results in a within-class evaluation of other students' prototypes or more often students simply asking family and friends for feedback. Although beneficial to an extent, the bias of this approach is self-evident.

9.3 Development of a New Model for Assessment and Feedback in HCI Courses

In the traditional sense, only a student and a lecturer are involved in the teaching and learning process. In this two-way process there should be a dialogic approach producing both feedback and feedforward responses (see Fig. 9.1).

Usually, this is the extent to which HCI, as well as any lectures containing HCI material, are taught. The knowledgeable lecturer is the authority on the subject area and is therefore able to both teach and assess the student, thus guiding the whole process from start to finish. In recent years, industry and recruiters have given a stronger message to academia, in that there is a need for more 'work ready' graduates. There has also been increasing attention given to 'impact' work, both at the postgraduate as well as research academic level, which involves academic achievement and outreach to the community and industry. In light of this, some teaching and research institutions are now involving industrial partners at the assessment criteria and evaluation stages (see Fig. 9.2).

While this is definitely a step in the right direction, and vital to produce 'work ready' graduates, there is still a lack of understanding from the students' in the way that users of a product perceive and use their applications. The criteria set by the lecturer and industrial stakeholders are lacking the true end-user perspective



Fig. 9.1 Basic lecturer knowledge exchange model (most widely used)

³<http://www.cardiff.ac.uk/software-academy>—Accessed December 2018.

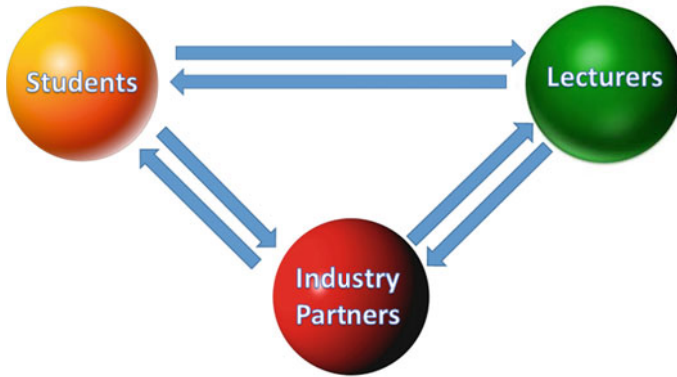


Fig. 9.2 ‘Advanced’ lecturer knowledge exchange model. What lecturers currently coinsider real world feedback within a curriculum

and feedback.⁴ For this reason, we are suggesting a model or framework which encompasses this further vital stakeholder; namely, the end-user (see Fig. 9.3). This involves a community of volunteers acting as part of an implicit macrotask solution

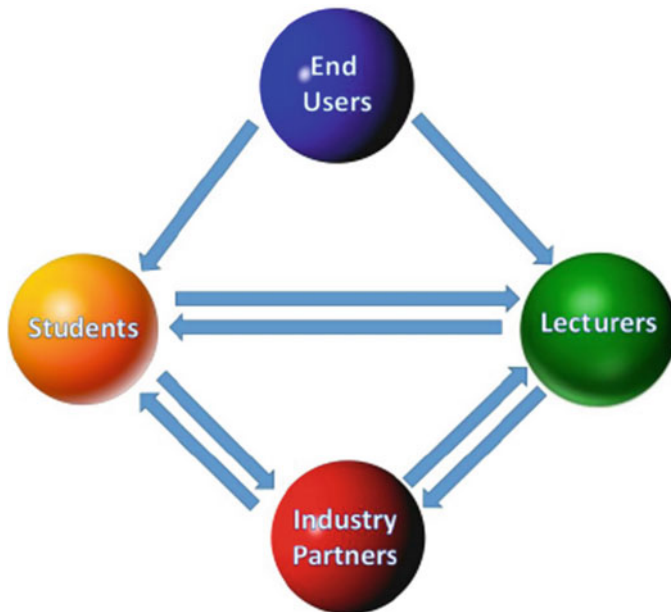


Fig. 9.3 Full real word lecturer knowledge exchange model, including all stakeholders such as end-users and their feedback

⁴Unless the end-user is actually the company or lecturer which sets the assignment.

involving crowdsourcing feedback from the community of users at the macro (or arguably micro) level (Cheng et al. 2015). Using this approach, a truly user-centred design approach can be enabled within an academic context.

There are, however, some risks attached to students receiving feedback directly from real-world users of their applications. Specifically, the type of comments posted on online forums on apps could be perceived as unhelpful or even ‘brutal’, inconsistent, given that users will have their own their criteria and personal reasons for their judgements, and may not have the depth and level of detail that students might expect from lecturing staff. It can also be argued that the receipt of comments from users directly, in isolation, is hard to classify as feedback. Specifically, given that the students will not have the opportunity to engage in further dialogue with the users who made the comments, they may therefore struggle to understand what the comments mean and will be less likely to utilise this data (Carless and Boud 2018).

For this reason, lecturers should go over the feedback with the students in order to contextualise them whether they are within the predefined rubric used for assessing the coursework within class or if the feedback covers areas different to the class assessment criteria. This method can help ensure greater consistency in the judgements made on the quality of the applications developed by the students, as well as help improve student assessment literacy, through which students will develop a shared understanding of assessment standards, of feedback comments, and of assessment processes (Price et al. 2012). Furthermore, this can create an opportunity for lecturers to review and reflect on the assessment criteria for students and to engage in reflection and dialogue with lecturers, which in turn offers the opportunity to improve student feedback literacy (Blei et al. 2003).

On this basis, and from the findings presented in the study which follows, we advocate in favour of crowdsourcing for the purposes of feedback providing a useful support for assessment.

9.4 Research Methods and Studies

In order to explore the potential of crowdsourcing feedback to augment existing feedback approaches, a pilot learning experience was designed and implemented with a small group of students in the summer of 2018. Alongside the pilot a research study was conducted using an exploratory case study approach. The aim was to explore the opportunities and constraints of crowdsourcing feedback, from the perspective of both students and lecturers. This allowed the researchers to remain open to emergent findings throughout the study, which would inevitably produce findings relevant to a range of stakeholders.

The first stage of the study involved three undergraduate students (two female and one male) who created Android mobile applications over a 4-week period and uploaded these to the Google Play Store. They were asked to publicise themselves and that they would within a few (8–12) weeks receive an invite for an interview to comment on the process and their experiences. After five weeks, feedback on the

students' apps left on the Google Play Store was collected and analysed in preparation for an interview to the students. Following this, the research team conducted semi-structured, discursive interviews with each student to understand how they responded to the feedback they were given, issues they faced and to gain their views on how the approach could be developed. As the research approach was exploratory, the interviews were analysed using the constant comparative approach. Without a priori codes, the researcher could remain open to emerging codes and themes within the data and across participants.

In stage two, the findings from the first stage were used to inform a series of small-group interviews with 16 undergraduate students in computer science and software engineering. This group was interviewed as they represented a key stakeholder group who were likely to gain feedback through this approach in the future, should the pilot be successful. Having had the context explained, the interviewers explored issues with the students about the potential benefits, the potential for harm, practical and ethical issues of the learning, assessment and feedback approach. Data analysis focused on coding and theming, informed by the codes generated in the first phase but remaining open to emergent codes.

Throughout the project, the research team kept observational notes and a reflective diary on their experience, detailing their experience as lecturers. This provided a valuable form of data for considering the operational constraints of the crowdsourcing feedback approach.

9.5 Deriving a Thematic Framework from the Experience

In this section, we present the findings of the constant comparative analysis of the pilot students' interviews under four categories which produce the basis for our thematic framework:

- (1) Design and Development Process;
- (2) Marketing;
- (3) Feedback;
- (4) Module Matters.

These are supplemented by findings from the analysis of the small-group interviews. The small-group interviews also revealed the potential benefits of engaging in this type of activity, from the perspective of the student, which is discussed in the final category:

- (5) Beyond Academia.

We discuss the details of these below.

9.5.1 Design and Development Process

The three students that participated in the pilot study were interviewed five weeks after first creating and uploading their apps. In this time, they had been expected to actively market their app and to have received feedback from users online. While this was not the focus of the research, when reflecting on their experience all three students began by describing their experience of creating, uploading and marketing their apps.

The design and development process was a key theme in which it was clear that each had different pre-conceptions and experiences of the app development process, as well as their own personal motivations for taking part in the pilot activity. While these students had self-selected to participate in the pilot and we might assume they would have an internal motivation to engage, the three students' engagement profiles were markedly different. One of the students stated that he only took part as 'something to do over the summer' and wasn't really interested in app development. When discussing the process of uploading and marketing the app he referred to his lack of knowledge on how to upload the app as a barrier (although detailed instructions had been provided). By comparison, one of the other students who described herself as 'creative' was clearly intrinsically motivated to take part, describing the process as 'easy' and 'enjoyable', although she had to research how to upload the app. Stating that she had 'learned a lot' from the process of designing, uploading and marketing, her new-found knowledge and experience motivated her to engage further '...I started making plans for my next app'. The second female student, although describing the overall process as 'fun', offered examples of how a lack of past-experience had led her to feel 'stumped' when designing the app. So, although she had a higher sense of motivation than the male student, initial difficulty designing the app became a barrier to engagement, as did her past-experience of social media when it came to marketing the app and receiving feedback.

9.5.2 Marketing

An important part of the process, and one that comes early on, is that of the students engaging in marketing their app. Although it was not designed to be a key feature of the learning experience, it surfaced as a key aspect; vital to receiving feedback. How the participants marketed their apps identified a range of influential factors regarding the platforms individual students choose to use.

Marketing the app differed in all three cases; as mentioned in the previous section, the second female student's lack of experience on social media became a barrier to her engagement. For her, marketing the app was a personal challenge, although she did market on Facebook and LinkedIn, she expressed a lack of confidence: "I don't really post on social media". By comparison, although the first female student had access to several social media platforms, she disliked the idea of marketing her app.

However, with an overall high-motivation to engage in the activity, she decided to research the best approaches. Consequently, she felt confident in marketing on her social media and even joined an ‘app developers’ Facebook group/page to reach an audience that she believed would gain her some ‘constructive’ feedback. The male student only used one social media platform to market his app (Discord). Instead most of his marketing was done through face-to-face interactions and one-to-one online private conversations: ‘I just passed it around people I knew...and ask if they could try and get their friends to download it...’.

It is clear from just the experience of the three participants that marketing approaches will vary. This will not only depend on an individual’s relationship with social media platforms but also personal attributes such as self-esteem and confidence. From the interviews with the students that participated in the pilot study, there was no clear consensus on whether having their name attached to the app may have influenced the positive feedback from their friends.

This was echoed in the small-group interviews. Using social media platforms for marketing the app was described by these students as having both positive and negative potential effects. It was highlighted that not everyone has an active relationship with social media, and one of the students argued that this could disadvantage those students ‘who don’t bother with social media’. Others expressed slight concern with their marketing abilities, giving similar responses regarding guidance and support of how to market effectively being incorporated into the module as well. However, for those who do use social media regularly, this was perceived to be an effective way to reach and widen the audience for an app—having family and friend support to offer feedback and share the app being one of the most beneficial in this case. It was also noted that the type of feedback you would expect to receive from those who are familiar with you may not generate honest feedback which could detract from the experience.

9.5.3 Feedback

By the time of the individual interviews, the students’ apps had been available to download for 5 weeks. In this relatively short time they had only received limited (under 15) comments each. All three students expressed some concern with how little feedback they had received, however all were optimistic that if further marketing via a University-wide email was employed, this would encourage further downloads and feedback. With the exception of one negative comment on the male student’s app, all comments received to date had been towards the positive (3 stars or higher).

When asked how they felt about receiving their feedback the first female student noted that she had received comments on Google Play and on the ‘app developers’ Facebook page and that ‘...it feels amazing, it’s really really good, but at the same time it is surprising because I don’t think the app is that good...’.

While positive feedback was generally welcomed, it was also clear that it was not necessarily helpful or what the students wanted. One-word feedback such as

'perfect' was described as '...not very useful... I guess it's like useful for people looking to download the app, but as a developer, it isn't really like feedback... because nothing is really perfect' (male student). Although the female students had only received positive feedback, they were keen to get constructive feedback. The first female student seemed quite open to any sort of feedback believing that even negative comments '...would just make it better...it would motivate me to go further'. However, the second female student considered the potential impact of negative comments on her: 'as someone who doesn't even post on social media I'm not used to having feedback like that of any sort... it would bother me slightly—if someone said horrible things without any constructive feedback—if given good reason then I would consider it, discuss it, try and improve for the future'.

While the male student initially demonstrated a level of apathy towards developing and marketing the app, when asked about how he felt if he would receive negative comments on his app he said '...I'm not bothered...uh because I probably wouldn't develop more apps to be honest...'; later in the interview he seemed motivated to address the negative comment he received, asking if he could update the app and change its name before doing any further marketing via the University-wide email. However, he also recognised that negative comments could be disheartening, describing the comment as 'harsh' although '...most of it is pretty true...'.

In the focus, small-group interviews, students were asked to imagine how they and their peers might respond to a range of feedback. An interesting consideration highlighted by some of them was the impact real-world feedback could have on a student's self-esteem; identifying potential issues with making the task a compulsory part of the module. Participants used words such as 'disheartening', 'detrimental' and 'demeaning'. However, it was also agreed that dealing with negative feedback is all part of the process and can prepare a student for the type of feedback they could face when working in the real world. A statement made by interviewee five argued that this would be the best way and best time to be faced with negative real-world feedback:

...there's a potential to receive some not nice comments but this is the world we live in and you can't release anything into the world without expecting some troll comments, so I think it's actually really good real world experience... if they can get that experience of dealing with real world commenters while they're in university, while they've got access to university support services... they're one of the best places to learn how to deal with that type of negative feedback...

During the small-group interviews there was substantial discussion about the types of feedback the participants would want or find useful. All the participants acknowledged the importance of having constructive feedback, even if it was from the real world and not from a lecturer. When asked how they would feel about receiving positive or negative feedback, their responses always reverted back to wanting constructive feedback of some sort, although eight students stated that receiving negative feedback would only encourage and motivate them to improve their apps. There were several references made to how this part of the task would be difficult to monitor, measure and regulate fairly, especially if the feedback was to be a factor in how the students were graded. It was agreed across all participants that other avenues would

need to be sought to make this part of the process fairer; participants offered ideas of how this could be achieved. Interestingly five of the interviewee’s suggested that the apps could be downloaded to University phones and then used to generate feedback from either university students or other members of the public such as school-aged children and passers-by. The idea of using other University students to judge the apps was mentioned on a few occasions, as well as using the Students’ Union as part of the marketing process. Participants were also asked how they felt about their identity being attached to the app and there were mixed views on whether that would be beneficial. Some participants believed that being anonymous could have a positive impact on the type of feedback received as well as if the feedback received was negative, it may not be interpreted as being a ‘personal attack’. However, other participants felt that having their name attached to the app might be beneficial at a later stage when writing their CV or developing further apps.

It is clear from the responses of the students who participated in the activity and those who were asked to comment on the potential of it for their course that it would be important for lecturers to support students as they receive feedback. This may include debriefing, helping them to interpret comments and identifying next steps.

9.5.4 *Module Matters*

In all interviews, students discussed the potential for this type of activity integrated into their courses within a module and as part of their assessment. While there were mixed views, all participants have similar ideas and opinions on how the task would benefit students and lecturer.

All three students that participated in the pilot agreed that this type of task would benefit any student doing a computer science or software engineering course. Having real-world feedback would complement the course ‘...as it gives you a different set of feedback that you wouldn’t get from lecturer’ (male student). However, there were varied responses regarding whether it should be a voluntary or compulsory part of the module. When asked if the feedback should impact on the student’s grade, there was some agreement that this may be difficult to ‘measure fairly’ due to individual students having differing levels of knowledge, understanding and ability. In the opinion of the male student, if feedback was to be a part of the grading then the module should be voluntary because ‘...I’m not entirely sure how useful it would be, I just don’t think there would be much feedback, I feel like it could be a waste of time... unless you can guarantee you’d get feedback’. However, the second female student gave a different response, ‘for people to be fully engaged with it there does need to be a grade attached to it...’.

Some of the students in the small-group interviews thought that making the module task compulsory would encourage students to familiarise themselves with the content of the module, while also giving them the opportunity to add an element of experience to their CV. Student 2 said it would ‘definitely be a good assessment’ task, explaining that he had already completed three years of computer science but

has yet to be given the opportunity to upload one of the apps he had written; much like another interviewee, he believes that being able to complete the full process of app development would be greatly beneficial. However, during discussion of whether this task should be made a compulsory or voluntary part of the module, participants raised several questions and concerns relating to how the real-world feedback would be generated and if used as a weighting on the student's grade, how it would be possible to make that fair. Participants asked how the feedback would be used to determine the student's grade, and the interviewers offered some scenarios to contextualise a potential process, for example positive feedback equalling a higher score or just getting feedback in general positively impacting the overall grade. From the discussions it became clear quite quickly in each interview that the participants were not very supportive of the idea of the feedback received being a grade predictor. One of the main concerns raised was the issue of reaching an audience that could guarantee downloads and 'constructive feedback'. While some participants acknowledged the importance of marketing their app, others did not initially take this into account until it was mentioned by the interviewer. However, there seemed to also be some concern of how marketing, in a particular way, may attract bias feedback and make the experience unfair if the feedback was to be used as part of the grading.

One of the ideas offered for overcoming the 'feedback grade barrier', was to include an element of grading but in the form of how the student responded to the feedback they receive. For example, if the students received negative or constructive criticism, they would then be required to show how they used the feedback to update or improve their app. Towards the end of the interviews, participants were asked if they felt that the lecturer should have some involvement in the feedback that each student receives.

9.5.5 Beyond Academia

The small-group interviews revealed a fifth category which did not emerge in the interviews with the three participants who uploaded their relevance of activity 'beyond academia'. This focuses on the relevance of crowdsourcing feedback for 'life after uni', gaining other perspectives on their work and applying their academic knowledge to the 'real world'.

There were several references made to life after uni, with students highlighting the need for real-world insight and experience as it would 'prepare [them] for working' (student 8). They perceived that gaining feedback from outside of academia would offer a 'non-bias view' (student 3) that judged them by usability of the app, rather than pre-defined marking criteria. Every participant agreed that they would benefit from taking part in the activity, as it would offer insight to 'what actual users think' (student 1), 'how usable [the app] is' (student 4) and what differences in feedback there are demographically, i.e. people from 'different socioeconomic backgrounds'.

In offering an insight into the sort of thought processes a student might have when designing and creating an app, they demonstrated a thought process which focused

on the user rather than a set of marking criteria: ‘...if I were the user what would I be expecting? What would I be wanting? And how easy is it to use the application?’ While others stated that ‘developing something that has an actual customer base’ would mean that they would need to ‘think from an outside perspective’, implementing the features the user would want ‘rather than the ones that you, [the student] would want...’.

Importantly though, students expressed a need to apply what they had learnt from their courses to this real-world context as it offers them the opportunity to experience the full life-cycle of writing, creating, developing and uploading an app: ‘...academic feedback is great when you’re trying to get an understanding of whether the process you went through is the right way... but whether it actually works or not is very much up to the population that you built it for rather than the academics’ (student 3).

9.6 A Discussion on Practical Challenges of the Findings and Scalability

In this section we discuss how our findings balance with practical challenges and how automation could provide opportunities to help alleviate these.

9.6.1 Lecturers’ Challenges

Several opportunities and challenges can arise surrounding the feasibility and logistics of adopting the process. Firstly, institutional rules and regulations need to be evaluated. This includes, but may not be limited to: intellectual property of student-created applications; ethical approval (given the human participation) and the logistics surrounding the purchase and management of external services (e.g. Google Play Store Developer Accounts). Clearly, there can be potential repercussions (e.g. GDPR⁵) for the representing institution with regards to how a student-written application is marketed and uses data, i.e. personal data or handles privacy implications within the product. Students will also need to understand these issues, which need to be supported by the wider degree programme, e.g. modules that address these points may not be delivered until the final year of study. Therefore the position of delivery relative to the process is an additional factor.

Building upon the feasibility, the second challenge area surrounds student uptake in the process. The outcome of the process above may dictate the release of an application to be an optional formative task alongside a degree programme or enable formal integration. For example, one potential stalling point is the inability to mandate that students set up accounts on external services. Additionally, from our experiences in recruitment, the number of students achieving a level which made them comfortable

⁵<https://eugdpr.org/>—accessed December 2019.

to submit the application online for feedback was low.⁶ This willingness is also exacerbated from the well-known fact that students tend to focus on targeting marks, and when there are no marks allocated to a submission very few will see the benefit of uploading for feedback. This may be understandable on the students' part, they may well be protecting their workload in doing so, however, a case can be made that the process of finishing a piece of work to the point that it is released publically provides utility irrespective of the feedback received.

The third challenge area surrounds handling the real-world feedback received as the quantity and quality of the feedback can be highly variable. This can enable or impede the project being useful for both students and assessors. Firstly, marketing applications can have unpredictable results in terms of the quantity received, which and could also be impeded or supported by institutional regulations and the design of external publishing services (e.g. the Google Play Store). Secondly, the feedback could contain damaging material and may not map to desired learning outcomes. This issue could be problematic for particular students or the cohort as whole; therefore, beyond institutional feasibility this risk is an important consideration to highlight.

Overall, this creates uncertainty for the assessor in wanting to ensure a consistent experience for the students, that could be challenging to resolve. One approach would be to moderate the feedback each student receives, but this would be time consuming and could be limited by the lack of useful feedback. This creates an argument against using this feedback for summative assessment is that the feedback received may not align with the assessment rubric. However, undertaking the process of releasing a product to market may bring sufficient utility in itself from a career development point of view.

Collectively, these challenge areas will create the feasibility for integrating this process formally into the design of a degree programme (or module). From a holistic standpoint, the sum of the above challenge areas could make the project infeasible over a short period of time (e.g. a semester), or feasible over a longer period of time (e.g. a degree programme). Therefore, we note the importance of considering the process as being formative, with its results enabling feedforward learning to the wider degree programme and into employment. From a module or degree programme delivery point of view, the students highlighted the need for explicit sessions to be dedicated to preparing a product for release and managing it over time, in addition to the development and technical processes. This could be an opportunity for cross-discipline influence in the learning experience students receive. We also encourage the use of automated tools to assist the lecturers in giving a concise representative feedback to the student, as well as insights to the lesson rubrics, based on the feedback comments. A prototype framework of this is seen in the next section of the chapter.

⁶This can sound as a negative connotation, reflecting low standards being taught, as well as a lack of confidence of the students in their abilities after completion of a module. There is however a realisation that all early stage developers (and all early stage professionals) have, which is that of lack of experience; therefore, an lecturer should distinguish whether there is a lack of confidence due to experience or material.

9.6.2 Towards an Automated Feedback Collection Process and Analysis

One of the limitations, identified by both students and lecturers, in the practical implementation of crowdsourcing for feedback from end-users is that of scalability. In other words, having a small number of students receiving a small amount of feedback from end-users is not a barrier to lecturers in reviewing the amount of information with the students. However, as the number of students taking part in this exercise increases, and/or feedback per student increases, there is a limitation in time available to scrutinise and make sense of these feedback comments. In order to address this issue, we would need to introduce automated systems, not to replace the lecturer and student within the feedback process, but as assistive tools to expedite the analysis stage of the feedback process. In this section, we share the process and software we have developed and are using in order to help students and lecturers understand and be able to use feedback faster. Although the software is written specifically to be used with the Google Play store, the principle and model is arguably externally valid to any textual feedback received using crowdsourcing. We are also able to use our tool in order to collate feedback from several developers' (students') comments and understand the feedback in its entirety representing a class rather than an individual.

Reviews for the applications on Google Play Store are a combination of a one to five 'star' rating system (one star being 'hated it' up to five stars for 'loved it') and unstructured open-ended comments left by users that downloaded the application. For students and lecturers, these reviews are an indication of the opinion of users about the application.

The star rating system will give an overall indication of the opinion of the user. The advantage of this type of rating is that the output is easy to analyse, for example, by computing the average or mode of all the reviews from the users and gaining a fast view of the general opinion about the application. However, it is a very basic feedback mechanism for a student or assessor, as it is missing the qualitative details which provide a richer understanding of the reasoning behind the star ratings. Conversely, open-ended comments are difficult to analyse, both from a human perspective and a computer perspective. From the human perspective, analysing and aggregating hundreds of open-ended comments can be very time consuming and, depending on the number of comments obtained, impossible to manually process by the lecturers and students involved in the project. In addition, open-ended comments can be harsh (and often explicit in nature), inconsistent (even within the same comment sentences can convey different sentiments and be contradictory to each other) or unhelpful (without valid reasoning or meaningful expression). But open-ended comments have very interesting properties and huge advantages such as carrying an important amount of knowledge about the users' needs and their opinions. This information can enhance the learning experience of the student, help them to improve their learning as they are richer and more comprehensive than a simple star rating, and can be more incisive than lecturer feedback. They can also help the lecturers to align their learning outcomes and assessment rubrics to the real world needs.

These textual properties are easy for a human being to both comprehend, as well as analyse. The only issue with human assimilation of the information is that of volume. In order to process large volumes of information, a human being would take a significant amount of time, and if the application proves to be successful, then it would become impossible for a single person to thoroughly be able to keep up with the incoming feedback. Hence, the need for computer assistance. Volume is not as much of an issue with software intervention. However, software in its current state, suffers from its own limitations; specifically, the opposite of humans'. Even if the field of research in opinion mining (Sun et al. 2017) is very active and more specifically in the context of app stores (Genc-Nayebi and Abran 2017) have produced promising research, the analysis of opinion mining in the context of app store feedback is still a very complex and challenging task to achieve. The reasons for this is human language, and its use within context. Feedback from users' comments are unstructured, very short in length (sometimes being a single word) and often composed of colloquialism and poor grammatical structure. In addition to this, most of the research presented above focuses on the identification, classification and summarisation of the comment to help the user to choose the best product when purchasing rather than more general sentiment. In our work, the main aim is to use these reviews as complementary feedback for the students.

In order to bridge the gap between human and computer assimilation and analysis of these comments at a large scale, we have developed a software base prototype to compliment the developers' process. We employed a user-centred design in which the requirements were taken from the user testing sessions presented in this chapter in order to produce the tool and student developers as well as lecturers to comment on the usefulness of the developed features of the tool. This is now at alpha stage and, although this is no the primary aim of the chapter, we present the model and framework of the prototype in order to inform the reader who may wish to use the same findings or method. We can also provide the software as open source upon contact at no charge; this tool is being developed constantly.

While we are developing an automated assistive tool, we primarily aim to enhance the user experience of the students or the lecturers when they must analyse the different kind of comments received. Beyond these important nonfunctional requirements, we developed two high-level sets of functional requirements for the prototype. The first functional requirement is to use automated sentiment analysis on the comments to provide an indication of their connotation (positive, negative, neutral). The second functional requirement is to be able to provide a more dedicated mapping of the lesson's rubric guidelines.

To successfully meet the functional criteria, the tool needs to firstly collect all the comments from an application on the store. In a second step, we then use Natural Language Processing (NLP) to split the comments up into sentences and automatically analyse the feedback and predicate the sentiment (Liu 2012) of each sentence and assign a positive, neutral or negative status to each. In a third step we cluster the different feedback comments according to the lesson rubric guidelines. Lastly, we implement a visual interface that will help the students and lecturers to avoid the possible overload of information. Below we describe the stages of the application at a

high level, almost framework approach. The reader is then able to use the best means and advances in algorithms at the time in order to reproduce effective software.

Step 1 (Scraping the data): The first step of this approach is to automatically collect the comments and the ratings. This step is relatively straightforward, and we collect the different metadata such as name of application, version, ratings, comments and dates of the posting. There is only one exception in that we choose not to store the replies to the comments by the developer. Only one comment by the author is permitted on the Google Play Store and no other replies by other visitors or the commenter themselves are permitted. We would therefore not gain anything from this information. The next phase of this step is cleaning the data gathered and to define which features could be useful for the learner and which for the lecturer. The cleaning of the data is still a very important and challenging part in any process involving natural language. App store comments are very short unstructured pieces of text that could contain colloquialism, often with poor grammar, bad syntax structure and rife with spelling mistakes which create difficulty in this process. There is also a further challenge in that these comments can contain fraudulent reviews (such as advertisements or spam) that could be difficult to distinguish from genuine reviews using artificial intelligence methods. Finally, an optional stage (and required by most teaching institutions) is the removal of unsuitable language. In this instance, care is needed to still, if possible, identify any constructive criticism, rather than disregarding the entire comment or sentence.

Step 2 (Manual Annotation): The goal of the second step is to create a ‘gold standard’ by which we will be able to assess the effectiveness of our automated NLP classification process. For this purpose, we need two corpora. A first, manual, annotation is performed on the first corpus. In order to do this, we manually annotated the polarity (Positive, Neutral and Negative) of the comments contained in the first extracted corpus. This will allow to evaluate the approach in terms of precision and recall (Buckland and Gey 1994). The manually annotated corpus contains different types of sentiment. Another element we can annotate is the correlation between the sentiment and the ratings. For example, a user could give a good general rating to the application but give very negative feedback on a specific part. We are also able to distinguish the topic of the comment, and whether it fits the predefined rubric from the lecturer; or, if it introduces new categories. In this step, we also manually extract the different topics from the assessment rubrics and create a terminology reference for our artificial intelligence engine to identify the different likely groups of assessment.

Step 3 (Classification): During this phase, a (trained) sentiment analysis is performed on each comment (at the sentence level) to identify positive, negative and neutral ratings. Each sentence will have a score between 1 (negative) and 3 (positive). We are currently using the Stanford CoreNLP natural language processing toolkit (Manning et al. 2014) to perform this task. In the second part of this step, we use a standard approach of topic modelling to discover the meaningful topics from each sentence; namely, the Latent Dirichlet allocation model (Blei et al. 2003). The annotations are

stored in a database along with the different metadata of the targeted application. The third part of this step is to match the topic discovered in the previous phase with the manual terminology created from the rubrics given by the lecturer, and classify the different sentences depending on their corresponding terminology. If the topic doesn't correspond to an existing cluster, the sentence will have 'unknown' for the topic and likely require further manual scrutiny.

Step 4 (Visualisation of the data): The result of the classification step is a visualisation of the percentage of positive, negative and neutral comment per rubrics classes or new classes. For the lecturer, this will be an ability to identify areas which are not covered in the lesson (or indeed any other lesson) and that require attention. It would also allow for a cross examination of their grading and a reflection to compare with real-world judgement. For the student, we are able to provide different level of granularity in what they need to address in their application. Effective visualisation means vary depending on the topic and rubrics, but what is important at this stage is the transition to different levels of granularity. At the high level, average scores and repeating phrases can be highlighted. When a developer wishes to investigate deeper, timelines (to detect update fixes) start to become more important. There is rarely a need to read each comment, as this would defy the point of the assistive tool, however, grouping of similar messages into a topic can be useful for identification of improvements to a rubric element.

Using the proposed model for the tool, we aim to assist the student developers to understand the feedback received from end-users faster and in a more concise way. This can also act as a mediator point for discussion with the lecturers. For the lecturers, the aim of this tool is to create an opportunity to engage with the students on the feedbacks they received, the need of the real world compared to the lesson plans and a way for them to review their own assessments' rubrics.

9.7 Conclusions and Future Work

This chapter has demonstrated how utilising public macrotasking for feedback on software-based student work has the potential to bring additional benefits for learners, through creating a process for authentic learning. To explore this in action, we have undertaken a pilot study that has emulated an assessment process involving the integration of public feedback as part of the formative feedback that a student receives. Overall, our method was well received by both lecturers and especially students, who were overwhelmingly positive about both the experience as well as the actual and perceived benefits.

From the perspective of students, we find overall support for bringing the real world into higher education through assessment as a means of supplementing student development. The crowdsourcing method employed in this study provided an opportunity to enhance the student experience, as well as to improve the richness of the feedback that they will receive on coursework. Specifically, in relation to

assessment design, it creates a more ‘authentic’ assessment experience; the need for real-world assessment experiences having been identified as an important strategy through which assessment can be linked with learning outcomes more explicitly and through which students can better demonstrate their skills and hands-on experience needed for the world of work (Wiggins 1990). Issues regarding the management of such an ambitious crowdsourcing task have been highlighted, such as the privacy implications of releasing prototype software by ‘amateur’ HCI practitioners, exposure and repercussions of the representing institution, suitability of crowdsourcing data gathered and communication challenges (Orsmond et al. 1996); especially for students. We also highlight the importance of initially associating a reward-based method as an incentive for students to use the macrotask crowdsourcing approach. We highlight that the main benefits to students included motivation to better themselves, a more diverse feedback experience, marketing and requirements of actual users and/or clients which are often different than the lecturer rubric.

Lecturers are also incentivised to use this macrotasking approach with benefits that often mirror the ones surfacing from the students, such as a richer experience and more feedback opportunity for the students (as well as to the lecturers themselves). We are also, however, given insight into the challenges of amalgamating the process and structure of academic assessment with real-world influence and ‘assessment’. Most notably, the legal and liability principles that a university adheres when software (such as mobile apps written by students) is perceived to be branded as belonging to that institution and the repercussion that may come from it. Care needs to be taken to ensure that students are exposed to the realities of development in a safe environment which uses criticism in a productive rather than a detrimental way.

Our experiences serve as the basis for future work in several areas. Building on the outcomes of this pilot study, the next stage is to implement this approach as part of a standard course. To achieve this, it will be necessary to integrate the development and release of Android applications as part of a module and its assessment activities, including the crowdsourcing of feedback. Further works could expand on this with whole cohorts and other types of products, not limited to the computer science domain. One example is the business studies field, or nursing, where macrotasking and crowdsourcing can be a rich way to increase outreach within feedback. Secondly, instead of a single semester and single module worth of feedback, we could expand the development of a single product over several semesters or produce consistent crowdsourcing macrotasks for feedback throughout the study period of a student’s academic life with positive effects. This may have the benefit of allowing students to develop their ability to respond to crowdsourced feedback effectively. Thirdly, the discussion suggests several areas for further, open tool development, to create an accessible learning environment that the wider education community can use to help implement the process in practice. This is represented by our suggestion for machine learning and text mining with natural language processing. Finally, it is our hope that this assessment and feedback approach can be implemented by readers with positive effects as well as having both ecological as well as external validity.

References

- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *The Journal of Machine Learning Research*, 3, 993–1022.
- Buckland, M., & Gey, F. (1994). The relationship between recall and precision. *Journal of the American Society for Information Science*, 45(1), 12–19.
- Carless, D., & Boud, D. (2018). The development of student feedback literacy: Enabling uptake of feedback. *Assessment and Evaluation in Higher Education*. <https://doi.org/10.1080/02602938.2018.1463354>.
- Cheng, J., Teevan, J., Iqbal, S. T., & Bernstein, M. S. (2015). Break it down: A comparison of macro- and microtasks. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI'15)* (pp. 4061–4064). ACM, New York, NY. <http://dx.doi.org/10.1145/2702123.2702146>.
- Genc-Nayebi, N., & Abran, A. (2017). A systematic literature review: Opinion mining studies from mobile app store user reviews. *Journal of Systems and Software*, 125, 207–219.
- Haas, D., Ansel, J., Gu, L., & Marcus, A. (2015). Argonaut: Macrotask crowdsourcing for complex data processing. *Proceedings of the VLDB Endowment*, 8(12), 1642–1653.
- Jessop, T., & Tomas, C. (2016). The implications of programme assessment patterns for student learning. *Assessment & Evaluation in Higher Education*. <https://doi.org/10.1080/02602938.2016.1217501>.
- Jonsson, A. (2014). Rubrics as a way of providing transparency in assessment. *Assessment & Evaluation in Higher Education*, 39(7), 840–852.
- Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5(1), 1–167.
- Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., & McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 55–60).
- Noy, N. F., & McGuinness, D. L. (2001). Ontology development 101: A guide to creating your first ontology.
- Orsmond, P., Merry, S., & Reiling, K. (1996). The importance of marking criteria in the use of peer assessment. *Assessment & Evaluation in Higher Education*, 21(3), 239–250.
- Price, M., et al. (2012). *Assessment literacy: The foundation for improving student learning* (p. 2012). Oxford: OCSLD.
- Soilemetzidis, I., Bennett, P., Buckley, A., Hillman, N., Stoakes, G. (2014). The HEPI – HEA student academic experience survey 2014. In *HEPI – HEA Spring Conference*, 21 May 2014 (pp. 1–40).
- Sun, S., Luo, C., & Chen, J. (2017). A review of natural language processing techniques for opinion mining systems. *Information Fusion*, 36, 10–25.
- Wiggins G. (1990) The case for authentic assessment. *Practical Assessment Research and Evaluation*, 2(2). <https://www.pareonline.net/getvn.asp?v=2&n=2>.

Chapter 10

The Mapping Crowd: Macrotask Crowdsourcing in Disaster Response



Ned Prutzer

Abstract Large-scale citizen involvement in disaster mapping is relatively recent yet impactful. Humanitarian OpenStreetMap (HOT) and Public Lab, two particular communities at the forefront of this shift, formed in response to the 2010 Haiti earthquake and the 2010 BP Oil Spill, respectively. This chapter compares and contrasts how these online communities employ crowdsourcing to aid in disaster response efforts. I employ OpenStreetMap (OSM) Analytics; Social Network Analysis; interviews with members from both mapping communities; my own experiences contributing via participant observation; and insights from OSM Users' Diaries and Public Lab research notes to do so. I also analyze community strategies and interface logistics involved in the work of both groups. Both communities are the result of ecologies of mobile applications, commercial imagery sets, government agencies, NGOs, and concerned citizens. The campaigns that result from these ecologies are branded as more efficient, cost-effective, and resonant with current political, economic, and social transformations. To help identify these changes, I overview imagined public uses of GPS explored within President Bill Clinton's and President George W. Bush's administrations. While the disaster response application fits within these intended uses on various levels, the scale of crowdsourcing applications demonstrated through these projects was largely unanticipated. In taking a historical perspective to the public use of GPS, I discuss how the rise of crowdsourced approaches correlates with an increased public unease with more traditional government responses to natural disasters that undergirds these communities.

10.1 Introduction

"Don't donate food or water or even money. Donate your time and skills to mapping Puerto Rico and help responders after #MariaPR," an organizer of a 2017 National Day of Civic Hacking mapathon tweeted. The mapathon supported Humanitarian

N. Prutzer (✉)

Communications and Media, Institute of Communications Research, University of Illinois at Urbana-Champaign, Urbana, USA

e-mail: prutzer2@illinois.edu

© Springer Nature Switzerland AG 2019

V.-J. Khan et al. (eds.), *Macrotask Crowdsourcing*,

Human-Computer Interaction Series, https://doi.org/10.1007/978-3-030-12334-5_10

253

OpenStreetMap's (HOT's) disaster mapping campaign to assist relief efforts in Puerto Rico. The National Day of Civic Hacking, an annual event that began in 2013 in partnership with Code for America, consists of different locally organized hackathons and camps to gather civic-minded designers and government actors in a given community with the intent of addressing the biggest problems the community faces through design.¹

The tweet encouraging HOT contribution in response to Maria valorizes small contributions of time in assisting the campaign over more traditional modes of help in recent decades like donating to the Red Cross. It recognizes a shift in views of disaster response in the wake of responses to Hurricane Katrina and Hurricane Maria. Public sentiment has grown increasingly distrustful of government and institutional disaster preparedness efforts in light of these events.

HOT facilitated scores of mapathons in response to Hurricane Maria and Hurricane Harvey. They made headlines over their ability to organize so quickly and effectively to fill such an urgent need.² Mapathons gather remote mappers as well as mappers local to the event to train and guide users through specific campaigns that need contributions. In such campaigns, HOT instantiates macrotask crowdsourcing to galvanize mappers across the globe to provide base layer data from aerial imagery. This data can help orient first responders when disasters strike. It can also have broader applicability depending on the campaign at hand, such as in support of government and NGO-led efforts to eradicate epidemics. More broadly, HOT's work includes remote mapping in developing areas around the world afflicted by Ebola, floods, volcanic eruptions, or other natural disasters.

Public Lab's work, also attuned to disaster response, is enacted differently—within close-knit community open calls; aerial imagery sorting activities facilitated largely through shared documents; and events that bring together Public Lab members, Gulf Coast organizations, and Gulf Coast community members. The community largely emerged out of academics, artisanal mapping experts, and concerned citizens gathering around the media blackout surrounding the 2010 BP Oil Spill. This work entailed kite and balloon mapping to capture its environmental impact and communicate its effects to the broader populace.

Members sorted aerial images taken from cameras attached to kites and balloon and “stitched” them on MapKnitter (Fig. 10.1), an open-source mapping platform where users place these higher resolution photos over satellite imagery or an OpenStreetMap (OSM) layer. Since then, Public Lab has experimented a great deal with modes of crowdsourced labor in sorting and analyzing open imagery provided by nonprofits and government agencies toward environmental monitoring.

In comparing crowdsourcing within these communities, I analyze how both models demand social interaction (be it virtually or in person) and collide with different technical and legal frameworks (be it from different algorithmic means of partitioning

¹National Day of Civic Hacking (2018).

²See Segal (2017), Yin (2017).

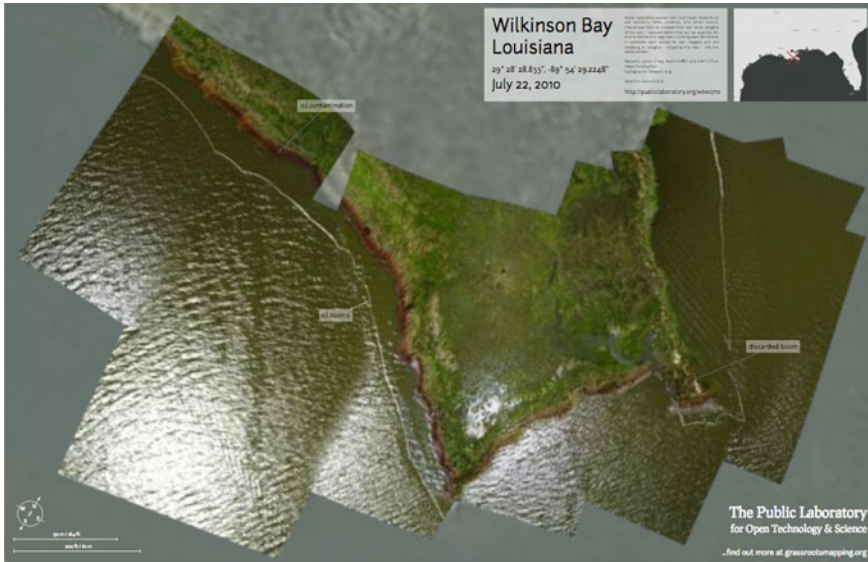


Fig. 10.1 An example of Public Lab’s image stitching via MapKnitter in response to the 2010 BP Oil Spill. The image depicts locations of oil spills and appears in the first issue of the Grassroots Mapping Forum (which now goes by the Community Science Forum), a Public Lab-authored journal

space, from different satellite providers’ imagery sets, or from different government agency datasets). Despite these larger formations, the act of mapping in these communities means paying attention to overlooked features in the environment and the available imagery, ones that reveal the complexity inherent in such work. Before detailing these dimensions, however, I define macrotasking in relation to these campaigns and explicate why these organizations invest in macrotasking as a technical solution.

10.2 Why Macrotasking?

Macrotasking has various distinctive features in comparison to microtasking platforms. While microtasking often involves route and standardized work, macrotasking campaigns must build in flexibility for important differences tasks may pose. Since projects in the latter are far less parsed out than in the former and often presented “as is,” interface and algorithmic designs that enhance usability and efficiency prove paramount. Macrotasking also requires subjective, and often multiple, levels of review.³

³Hass et al. (2015, pp. 1643–1644).

What qualifies as macrotasked work is not an exact science despite these general criteria. HOT's work, while fairly parsed out in how it organizes mapping tasks for a campaign, depends on various usability and technical structures to facilitate complex work that requires a longer engagement per task than most microtasked campaigns. While there is a clear strength in having a consistent cohort of mappers contributing, turnover in those contributing is constant—an affinity such work shares with microtasking communities.⁴ The design particulars in these macrotask campaigns work to minimize turnover to ensure a dedicated contingent of mappers persists, be it through the ease of use or modes of gamification.

Public Lab's work is not nearly as parsed out and often does not have structures of review built into the interface itself. Instead, methods of review often come from formalized events, or from mappers reaching out for help on the various modes of communication its website supports. Further, it is a more intimate approach than a far-reaching microtasking solution would provide. For Public Lab efforts, the focus is always on affected communities having the ability to advocate for themselves, rather than leaving the fate of their problems solely to "the crowd". The merits and dimensions of macrotasking for these groups are thus often highly situated.

While there are several unique facets to this work, its inherent urgency certainly stands out as a common issue. If mapping a given area devastated by a natural disaster and lacking OSM base layer data prior to the crisis were to take too long, the utility of it for first responders would be lost. There would of course still be use for the campaign if the area is one that is slated to be hit by further disasters as a result of climate change, with the Gulf Coast being one example. Even in such cases, data may need additional double checking through another later mapping campaign to make sure the base layer is as up to date as possible to assist in the response properly. In Public Lab's case, the longer it takes to identify and document environmental concerns, the more detrimental they can be for the impacted area.

The utility of publicly available GPS data and aerial imagery to foreground citizen-led interventions into environmental monitoring in such cases was largely unanticipated. Leading up to the Clinton administration's decision to lift selective availability of GPS to open the system up to public use in 2000, the administration posited a range of potential benefits to doing so, both internally and to the public at large. Such imagined applications pitted GPS, for instance, as a modernization project toward "the new economy;" as an application toward locating vital infrastructure toward disaster planning; and as part of assemblages toward public safety and monitoring criminals.⁵

⁴See Jemielniak (2014, p. 13).

⁵Richard Armitage hosts Ask the White House (2004), Bryan (2016), Charles G. Groat hosts Ask the White House (2004); "Fact Sheet: U.S.-EU Summit: Agreement on GPS-Galileo Cooperation," *George W. Bush White House Archives*, accessed June 26, 2004, <https://georgewbush-whitehouse.archives.gov/news/releases/2004/06/20040626-8.html>; "H.R. 2561—Department of Defense Appropriations Bill, FY 2000," *George W. Bush White House Archives*, July 21, 1999, <https://georgewbush-whitehouse.archives.gov/omb/legislative/sap/106-1/HR2561-r.html>; Press Briefing by FEMA Director David Paulison (2006), Remarks by President George Bush, Prime Minister of Ireland Bertie Ahern, and President of the European Commission Romano Prodi in Press

Yet, perhaps most pertinently, the value of public-oriented GPS for macrotasking fits within the broader discourse of “reinvention” the Clinton and George W. Bush agencies framed GPS investments within, especially within the former’s Reinventing Government program and the latter’s E-Government initiative. Reinventing Government intended, in large part through collaboration across government agencies, “to reform and streamline the way the federal government works.” It sought a government that, in the words of Vice President Al Gore, “works better, costs less, and gets results Americans care about.” It also accorded with broader Clinton policies that sought to replace government expense with efficiency, and to move away from “entitlement” among the citizenry toward “empowerment.”

Meanwhile, the E-Government Act in 2002 established “a broad framework of measures that require using Internet-based information technology to enhance citizen access to Government information and services.” This goal of broadening access to information is tied in part to GPS use.⁶ These themes within opening up GPS of ensuring efficiency and activating the citizenry toward productive ends manifests within the work this chapter examines.

In the next section, I begin to untangle these complexities via partnerships and technical dimensions of HOT campaigns before doing the same with Public Lab initiatives. A group of self-organized OSM members launched HOT in response to the 2010 Haiti earthquake. Since then, HOT has galvanized various humanitarian-based mapping campaigns in conjunction with state and NGO actors.⁷ But HOT is only one node in an emerging humanitarian network dedicated to putting contemporary platform innovations to use.

10.3 Humanitarian OpenStreetMap (HOT)

HOT’s partners include the American Red Cross; The Bill and Melinda Gates Foundation; the Digital Humanitarian Network; the Global Facility for Disaster Reduction and Recovery (GFDRR); the Knight Foundation (which has also supported Public Lab); the Humanitarian Innovation Fund; the Peace Corps; the World Bank; and the US Department of State. The latter is involved both through its Bureau of Population, Refugees, and Migration that assists, for instance, in HOT Uganda and Turkey refugee mapping projects and through its MapGive training program for volunteer mappers.⁸ Government agencies thus not only often have reason to partner with such mapping work, but also actively invest in programs that promote it for enhanced

Availability (2004), President Discusses War on Terror at Naval Academy Commencement (2005), U.S.—Canada Smart Border/30 Point Action Plan Update (2002).

⁶For more on these initiatives, see Kamensky (2001), PUBLIC LAW 107-347—DEC. 17 2002 (2017), and The President’s New Freedom Initiative: The 2007 Progress Report (2017). For a media studies perspective on how Reinventing Government’s view on welfare specifically sublimates into media forms, see Ouellette and Hay (2008).

⁷Warren (2010, p. 39).

⁸Partnerships (2017).

understanding of different environments. Other partners include the Humanitarian Data Exchange (HDX), the United Nations Office for the Coordination of Humanitarian Affairs (OCHA), MapBox, DigitalGlobe (the leading satellite image provider), and RadiantEarth (which “offers solutions to fully realizing the potential of earth observation for positive, even life-changing global impact”).⁹

To better understand the practices, technologies, and interactions of HOT contributors, I contributed to projects that include, but are not limited to, malaria mapping for Cambodia, Senegal, Rwanda, Angola, and Laos and disaster mapping projects addressing flooding in Japan, wildfires in South Africa, and areas affected by Hurricanes Harvey and Irma. Much of what follows relay my experiences with a pilot HOT internship program as a Malaria Mapping intern on a set of disaster mapping projects with the Clinton Health Access Initiative (CHAI) as the requesting organization. Such requests are needed before OSM directs campaigns. The area often must lack a local community for the impacted area that is able to map for themselves in a locally informed way from the outset to warrant such a campaign being formalized.

10.3.1 Characteristics of HOT Disaster Campaigns

Mapathons, as previously mentioned, are a prime component of HOT campaigns. Missing Maps runs many of HOT’s mapathons. Supported by global humanitarian efforts and perhaps most notably by the Red Cross, Missing Maps is dedicated to, in its own words, “putting the world’s vulnerable people on the map.” It sponsors a great deal of mapathons across the world and sustains a workflow of gathering data from remote mappers, enrolling and training local mappers to ensure data quality, maintaining continuous data contribution moving forward that can be locally informed, and partnering with NGOs to use that data toward development goals.¹⁰

Aside from mapathons, most users learn how to contribute via online tutorials that groups like Missing Maps and dedicated mappers produce and feedback gained from comments on their mapping. Given the lack of face-to-face interaction in these cases, the technical structure of HOT campaigns is crucial. Aside from the logistics involved to ensure as much ease in mapping as possible, the task instructions, comments, and conversations involved in validation are all critical characteristics of the campaigns (Fig. 10.2).

Task instructions often narrow down the points of emphasis for the task. Those in charge of HOT tasks are very aware it is easier for mappers to enter into a task with the knowledge they are only mapping roads or only mapping buildings, not both. The recognition of the cognitive load involved further signals how HOT factors such considerations in as much as possible.

⁹See *Radiant Earth* (2018).

¹⁰Missing Maps (2018).

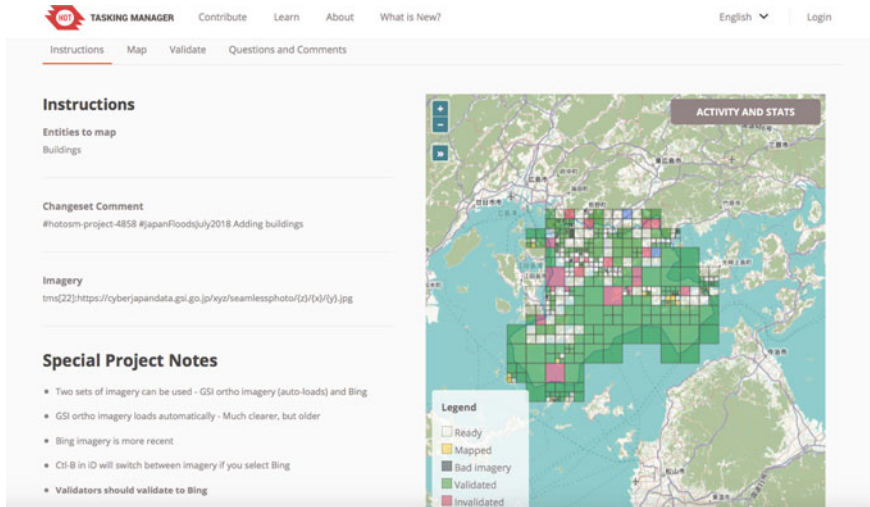


Fig. 10.2 A screenshot of a task on the HOT Tasking Manager. Instructions specific to the task at hand are on the left side of the screen; on the right lies a parceled-out map of the area being mapped that is color coded to indicate which boxes are complete, in need of validation, or in need of mapping

The design of the maps HOT uses to divide tasks within campaigns equally demonstrates attention to these concerns. Mapping of different buildings and infrastructures occurs by parceled-out tasks on a map that are assigned numbers. Campaigns themselves are categorized in terms of which require expert mappers and which do not so users can select ones they feel match their level of expertise. Nevertheless, the work of these campaigns demands social interaction across these levels of expertise. At least two mappers are necessary in order for a block within a task to be considered complete: one to fill out the block and mark it as complete, and another to validate that the work on the block has been done correctly.

As one maps, communication with mappers working on the same task occurs through changesets and the Task Manager comments. Numbers are incorporated within changesets to convey how many mappers have contributed to a given campaign, which also gets designated within the changeset comment via the task number. Mappers are then encouraged to add what they actually did (“added buildings,” “modified highways,” etc.) and to save their work often, so that if work must be reverted for any reason it is far easier and faster to accomplish. While Task Manager comments are more for comments between mappers and validators, changesets are more global to the rest of the OSM community. In the event of a future project in the area, they can signal why features may have been traced when they may not show up in a given image set being used for the current initiative.

Members are also often able to contribute to how the areas in a given task are parceled out toward efficiency and usability prior to task initiation. MapSwipe is a mobile app HOT developed to determine which areas should be included on a given

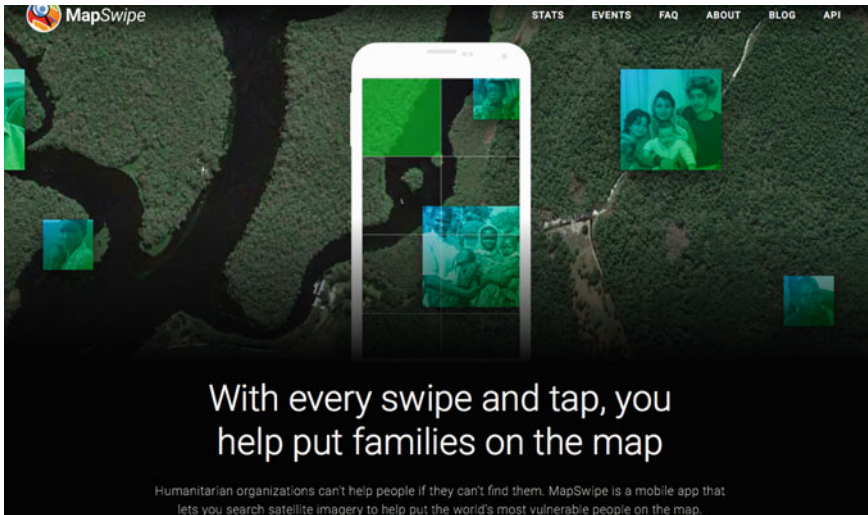


Fig. 10.3 The home page of MapSwipe, with an image of a phone screen that mimics what the MapSwipe interface looks like

project map. Users swipe through images from areas of focus in different campaigns and indicate if there are, are not, or maybe buildings in the images shown. Whereas other approaches have led to small squares with little, if not nothing, to map, a HOT employee claimed that the “funky” algorithm helps create more irregular shapes that would make mappers feel like they were contributing by including a decent amount to map. It also does not overwhelm them with so much that they would move on from that particular tile. The aim was to strike this balance so mappers would feel inclined to map more than one task each time they logged on. HOT knows if tiles take 10–20 min to complete, this is much more likely than if they take half an hour or more (Fig. 10.3).

Efficiency is of course of utmost concern, and ultimately informs decisions in what platforms and tools to use in disaster mapping projects. The editor typically used by beginners, iD editor, is a platform that is easier to map with but harder to square as accurately or as quickly as JOSM. On the iD editor, widely perceived as better for beginner mappers to use, the denser a tile is in terms of buildings, the less and more selectively it decides to display them when one zooms out to try to see building differences better.

JOSM carries various different preset tools that enable quicker mapping. One for buildings, for instance, allows a mapper to draw a line on one side of a building and then drag their mouse to the opposite side of the building to create a rectangle, automatically tagged as a building. To map a round building like a hut, one simply makes a triangle of three points at its border, hits a keyboard shortcut that renders it a circle, and copies and pastes from there over others, provided surrounding huts are roughly the same size.

On both iD and JOSM, a mapper can map quickly by copying and pasting a building shape to superimpose on similar building shapes. But on iD, when tiles present buildings of myriad sizes, one needs to first place nodes connected by lines for the four corners of the building, then additionally scroll to the Edit Features menu to tag it as a building manually. The JOSM buildings preset does so automatically.

HOT has a reputation for being welcoming in theory but terse in practice within this workflow. HOT's user guides vouch that these tendencies should not dissuade mappers, particularly beginners, from continuing on and completing their intended work.¹¹ For one Cambodia task I worked on, I received a message through the site entitled "Task #836 invalidated," with the curt message, "Missing buildings." The user left a similar message on another tile shortly thereafter, without any real suggestions for further resources to refine my skills going through the imagery or any constructive feedback. To compare what HOT encourages to validators' responses, here is an example of a comment I received from a fellow intern following a webinar on best practices in validating on some of my earliest mapping work while interning that occurred several weeks beforehand: "Thanks for your contribution! Make sure to square the edges of your buildings. If you are using iD editor press "S" or if using JOSM press "Q" Thanks again!" Had the comments I received before been in this style, I would have likely improved my mapping much more quickly.

Task completion can thus take quite some time depending on the project at hand. Part of the reason, depending on the task, can be due to misunderstandings or community debates on certain standards and classifications. Road classifications, for instance, often prove particularly contentious. In light of these difficulties, as part of my experiences collaborating on mapping Uganda in response to the refugee crisis, I participated in a webinar with a representative on the ground to receive explanations about the basics of different road designations in the area. This work helps develop algorithms for the United Nations High Commissioner for Refugees toward analyzing the spatial distribution of WASH (water, sanitation, and hygiene) facilities. HOT in Uganda also organizes mapathons to provide further data to NGOs and refugees on the ground.¹²

The webinar demonstrated how one can tell the differences in what the road looks like, what it runs through, and what its larger role is in the road network being mapped. The purpose of the webinar was to give HOT interns the opportunity to speak with a Ugandan HOT worker to know the correct road designations to use and how they would appear on aerial images. Though contributing may have been more difficult for a community member without this opportunity, task instructions still delimited two types of road designations appearing in the task. Outside of these internship experiences, task instructions available to all mappers working on a given campaign provide much of the same information. These instructions often narrow down the potential road classifications one is likely to come across in a given task to two or three potential types of roads.

¹¹Ibid.

¹²Urban Innovations: Crowdsourcing Non-Camp Refugee Data (2017).

Through the webinar, however, HOT interns were able to gain far more on-the-ground knowledge in a more immersive way that made mapping far less complicated. The internship program modeled mapping practices and behaviors to make the campaigns and their output work as best as possible. The next section delves into the background of the interns and the internship program, available statistics relevant to the program and other malaria eradication tasks, and the challenges and interfaces involved in HOT's work more broadly.

10.3.2 Case Study: Malaria Mapping Internship Program

Interns ranged in their OSM experience. Several came from the YouthMappers program, which “supports university efforts to offer meaningful global learning experiences, build a socially engaged citizenry, enhance long-term scientific capacity around the world, and foster youth leadership.”¹³ The implication of these modes of self-organizing on citizen participation is thus overt in the missions of many of the groups associated with this labor.

The YouthMappers had several years of experience mapping on the ground with OSM. Other interns were North American students (many women) studying GIS with limited OSM experience. One notable intern was a middle-aged worker looking for a career change and seeking to gain experience with humanitarian data projects.

HOT's internship program was largely unprecedented for the organization. HOT had individualized internship experiences such as in social media outreach in the past, but never an “army of interns,” as one HOT employee described. The need for the Eliminate Malaria mapping internships arose from HOT being off track on its mapping goals for the South Asia region. HOT was unable to take in all of the 160 applications they received—far more than predicted. Though HOT mulled accepting only 20 of the applicants for a more tailored experience, the program took in 60, without daily mapping check-ins in favor of weekly ones via email.

There was a weekly quota of 2000 buildings to map per week. This quota could vary if a given intern was placed on validation for a project, or if an intern was put on a project involving road mapping, it would change to an agreed upon quota in kilometers mapped. Regardless, the expectation was that mappers would devote about 12 h a week to the assigned project in meeting this quota. I went far over in reaching the quota initially as I adjusted to Java OpenStreetMap (JOSM), an advanced interface through which one can edit and publish OSM data. Other OSM editor interfaces include iD (which most, myself included, use as a beginner) and Potlatch.

¹³About Us (2018).

Unlike other projects, where the data might also be for purposes of navigation, the Eliminate Malaria campaign was a strict building count so that workers on the ground could know how much spray and how many bed nets should be packed in visiting a given area. Below I list statistics for a sample of these projects, determined by those that had statistics available via OSM Analytics. OSM Analytics is a site with comparative data that can display the degree to which the amount mapped in a given area may increase over time, as well as visualize the impact of these projects (Figs. 10.4 and 10.5).

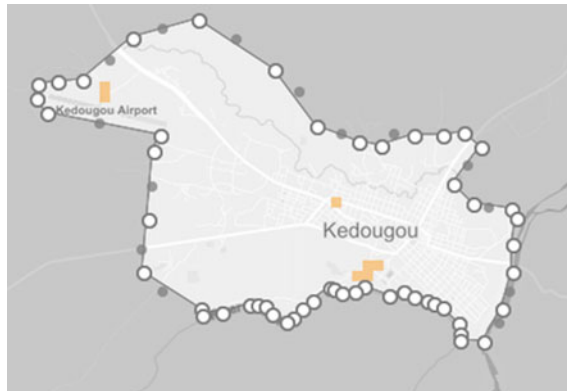


Fig. 10.4 A rendering of OSM data available in 2014 from OSM Analytics of the area mapped in Task 3327—Malaria Health Map—Kedougou, Senegal. By then, OSM members had contributed nine buildings to the area

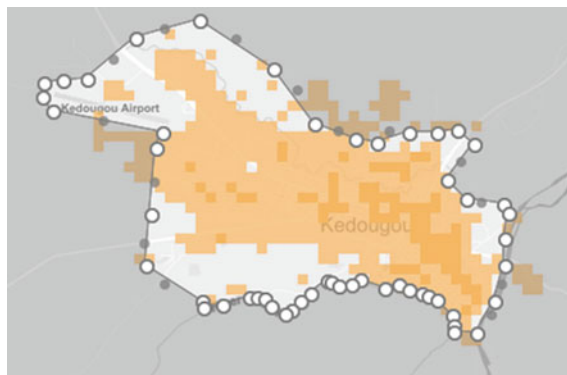


Fig. 10.5 A rendering of OSM data available in 2018 from OSM Analytics of the area mapped in Task 3327—Malaria Health Map—Kedougou, Senegal. By then, OSM members had contributed 11,950 buildings, in larger part due to three different HOT tasks

10.4 2014 and 2018 Building Count Comparisons in HOT Malaria Mapping Campaigns¹⁴

Task number	Country	Organization	Percent mapped (%)	Percent validated (%)	Before task	After task
2136	Guatemala	CHAI	94	92	0	21,009
2166	Guatemala	CHAI	100	100	0	36,953
3327	Senegal	N/A	100	93	9	12,546
3979	Senegal	N/A	100	100	0	480
3980	Senegal	N/A	98	100	0	1,484
3981	Senegal	N/A	100	100	0	190
4168	Mali	MSF	82	2	2,527	20,198
4265	Mozambique	Peace Corps	100	39	0	5,444
4304	Mali	MSF	99	4	16	22,742
4305	Mali	MSF	100	3	851	3,661
4317	Botswana	CHAI	100	100	114	24,599
4338	Botswana	CHAI	100	100	0	2,261
4339	Botswana	CHAI	98	100	9	10,258
4340	Botswana	CHAI	100	100	0	3,789
4341	Botswana	CHAI	100	100	0	2,032
4382	Botswana	CHAI	100	100	1	9,365
4425	Mali	MSF	100	2	1,346	21,018
4433	DRC	MSF	100	47	0	2,416
4439	Mali	MSF	75	3	38	21,146
4633	Papua New Guinea	MSF	99	8	0	18,665
4746	Mozambique	Peace Corps	97	76	7	45,877
4762	Mozambique	Peace Corps	99	30	0	9,348

As the data suggests, HOT campaigns serve as an impetus to map areas that barely have a footprint, if any, on OpenStreetMap (or any other mapping platform). The new base layer data that results can then be available for any future aid campaigns that may be needed in those areas.

But many challenges can arise in this work that throws its accuracy into question. While tracing building footprints as accurately as possible is ideal to estimate the location, shape, and size of the building for a number of occupants, what mappers ultimately have to trace through the aerial view is instead the “roofprint.” In assessing the damage, as is the point of many HOT tasks, these prints can be misleading. An

¹⁴These statistics are up to date as of September 25, 2018.

otherwise normally standing building could lack a roof, and more importantly, a building that has a roof could very well be otherwise dilapidated.

Another challenge to accurate mapping in disaster mapping projects is the recency of the available aerial imagery. Usually, the mapper does not have this information available in assessing different imagery sets (be it from Bing, DigitalGlobe, or another provider) to use. In making judgment calls about which imagery set to use, one relies on aspects of the environment—such as trees and roads—to “speak” to its recency. If trees are larger, more numerous, or absent, or if roads appear wider or straighter, those changes (often man-made) in the landscape can point out if one imagery set should be used over another.

The trouble, again, is that this is not an exact science. The changes I spoke of before can aid mappers, but again, there is often no means during the act of mapping to tell when sets of aerial images were captured. At times, task instructions stipulate which imagery set to use out of recency. With various other tasks, it is a judgment call, with HOT representatives identifying such calls as equal parts art and science. The archive of available images and the subjective calls mappers must rely on always calls into question the notion of any map being “complete.” These situated features of mapping campaigns thus go far beyond what is considered as microtask labor conventionally.

Roads are another element of the natural environment that can provide helpful information in deciphering aerial imagery. They can distinguish between trees and huts, which can look similar in aerial images. If a road ends and a circle of such shapes appears, the road likely ends in the middle of a hamlet, indicating that what is shown are huts rather than trees. If such a shape appears in the middle of the road, it is unlikely that a building would be placed there. What is pictured is likely tree cover over a road. In spite of the grand view aerial imagery often affords, it is the mundane yet important objects tied to different spaces that often factor in most meaningfully.

To further help distinguish huts from forestry, HOT often encourages imagery offsets. Typically, on OSM, one could check imagery alignments with GPS traces. If a trace of one walking down a road strays from the aerial image of the road, that indicates the aerial image at hand from a given imagery set must be repositioned to match how data is being drawn using other imagery sets. One can do this by dragging the background on iD or JOSM. For the areas that HOT projects focus on, such differences can prove significant. Even a difference of a few years between imagery sets can mean a great deal in terms of accuracy when mapping areas often susceptible to flooding, for instance.

Legal and technical dimensions behind imagery acquisitions and interface operations are worth noting. For many projects, one must consent to agreements over the use of the imagery the project borrows. As one example, a NextView license agreement I signed for one such project, one that would also come up in post-hurricane imagery I would later work with, stated I would not use the imagery outside of “digitizing OpenStreetMap data for humanitarian purposes.” The copyright for the imagery, in turn, constantly turned up in the imagery itself as I was mapping, a constant reminder of the political economy of these imagery sets.

Innovations in high-revisit satellite imaging to capture aerial images of an area for monitoring human and nonhuman activity scores of times daily are occurring within a very small market dominated by mergers between select imaging companies and communications firms. DigitalGlobe, one such company, owns over half of the market.¹⁵ The work of these communities thus responds to the closed nature of aerial imagery and spatial data platforms, rendering communities of interest passive instead of active agents in issues that affect them.

When it first became involved in such initiatives several years ago, DigitalGlobe's role within these collaborations was largely unformalized and on a case by case basis. Early in 2017, however, it established a formal procedure via its Open Data Program that it would make before and after imagery of areas affected by natural disasters publicly available based on set criteria. Like HOT, their protocol includes major disasters in developing countries that a committee within the program decides mandates open imagery. Public Lab, in contrast, is a more public-facing process, with the community itself deciding what kinds of partnerships and initiatives seem most appropriate to the situation at hand.

Imagery for use in these campaigns is often made available under a CC-BY SA 4.0 license. ODbL, the Open Database License OSM data is made available for use under, is not fully compatible with the CC-BY SA 4.0 license. But with companies' formal acknowledgment of indirect credit on the OSM Contributors page and their waiver of a component of the license that proscribes "downstream restrictions" to access to the data in question to OSM and its users, members can use the data without restriction.¹⁶ The latter essentially permits the digitization of the data in question on OSM (as in, for instance, tracing over imagery) and its redistribution on the platform under ODbL given OSM's status as a "living" map.

Though HOT and Public Lab are similar in their open-source knowledge production and the nature of their collaborations, their approaches can diverge. As one Public Lab member noted to me, the difference in primary goals between "authoritative data" for a requesting organization and "community autonomy" result in different community models. The "dream" this user expressed would have two dedicated groups documenting a large-scale event—one that would survey on the ground and generate their own mapping images (rather than having projects rely on proprietary imagery sets with unknown dates), and another that would be well-versed in stitching remotely.

The user's expressed desired is a commentary on MapKnitter's design and the balance between what aspects of stitching get automated and what gets left in users' control. The more left onto the user in these responsibilities, the more labor is involved. That proves particularly taxing if one's accumulated images provide a large sample to work with. It is not unusual to have hundreds or thousands of images captured within a given trip. The next section details several Public Lab interfaces and technologies

¹⁵Scoles (2017).

¹⁶"Use of CC BY 4.0 licensed data in OpenStreetMap," *OpenStreetMap Blog*, accessed August 13, 2018, <https://blog.openstreetmap.org/2017/03/17/use-of-cc-by-data/>. See also Creative Commons Attribution 4.0 International Public License (2018).

before overviewing a particular disaster mapping event in response to Hurricane Harvey.

10.5 Public Lab

Public Lab's cartographic work typically does not fall within a conventional crowdsourcing model. It is smaller in membership than efforts like OSM and HOT and better known for community-driven designs of low-cost data capturing technologies so that public can advocate for themselves in capturing spaces of everyday life. Public Lab's website sells affordable kits for mapping that include 1000 ft spools of string, gloves, carabiners for camera rigs, rubber bands, and kites, mylar balloons, or even poles for aerial imaging. Members employ these means of capture to illuminate the dynamics of their local communities. They can purchase balloon or kite mapping kits from Public Lab's website. Proceeds from kits go right back into funding Public Lab.

One notable Public Lab project is Ann Chen's maps of False Creek, "a narrow inlet bordering downtown Vancouver." Chen uses balloon mapping to chart the multiplicity of histories inherent in the site as well as its shifting shores due to changes in human activity. Specifically, Chen traces how alterations in economic production tied to the land, spanning from indigenous hunting and fishing to industrialization centuries later, had material and spatial consequences, with False Creek now left less spacious and narrower. This is on one hand due to landfills from industrial development, but it is also due to postindustrial measures to attempt curbing the problem by building new structures atop pilings, under which creek water has now collected.¹⁷

Public Lab's workflow is in many respects adhoc; it reflects the group's beliefs in critical data consumption and collaboration, rather than traditional specialization, as the key to innovation.¹⁸ Public Lab has several modes of communication to gain feedback on work in spite of how distributed the community is. Aside from notes and comments on the site and email lists, Public Lab hosts OpenHour Zoom calls monthly. They vary in terms of their topic. Past OpenHours, which have been running since 2014, have largely reflected the content of Public Lab research notes. The OpenHour topics have ranged from mapping, air quality monitoring, thermal imaging, and hydrogen sulfide monitoring. In the spirit of transparency, Public Lab makes call notes available on its website.

¹⁷See Ann Chen, "Balloon mapping False Creek," *National Geographic blog*, last modified March 30, 2015, <https://blog.nationalgeographic.org/2015/03/30/balloon-mapping-false-creek/> and Ann Chen, "False Creek, Vancouver, January 2015," *MapKnitter*, accessed June 4, 2018, <https://mapknitter.org/maps/false-creek-vancouver-january-2015>.

¹⁸See Dolan (2010, pp. 33–50). I find it worth mentioning here that, as Evgeny Morozov does with digital mapping, Dolan points out that the emergence of digital adhocacies can have both positive and negative valences, such as with hate groups. For a condensed take on Morozov's connections in this regard, see Warren, "Grassroots Mapping," p. 24.

Public Lab is currently supported by various foundations and funding agencies (including Google, NSF, the Posner Foundation of Pittsburgh, the Gordon and Betty Moore Foundation, and the 11th Hour Project) as well as a range of in-kind donations (including server space from the MIT Media Lab and Rackspace).¹⁹ Prior supporting organizations include the Knight Foundation, the EPA, Microsoft, MapBox, Mozilla, Development SEED, and the American Anthropological Association.²⁰

While many of the most significant MapKnitter projects like Chen's focus on pollution and waste, several have focused on matters of disaster management and community-led stormwater monitoring. Applications range from capturing aerial imagery of annual flooding in Jakarta to survey images toward mapping stormwater runoff in New Orleans.²¹ Both projects have made use of the primary method of publishing work on the Public Lab website: research notes.

When one reviews member contributions on research notes, the close-knit and highly concentrated nature of the community becomes apparent. While about one in three notes receives no comments, in those that receive two or more (about one in five), eight users emerge as key cogs in the site's dialogue. They represent less than a tenth of the total contributors over a 6-month span of notes I reviewed.

Many of those select users serve in formal roles within Public Lab. In the network, it is clear that one user, demarcated in the network visualization below through the node of deepest red, pens or is present in a great deal of the online interaction (Fig. 10.6).²² New users often enter into Public Lab's work through hyperlocal projects, be it through initiatives like Chen's or imagery sorting workshops like those which emerged in Harvey's aftermath to map oil spills and flares.

10.5.1 Case Study: Hurricane Harvey Disaster Mapping

Leading Public Lab figures and concerned members quickly self-organized in Harvey's aftermath. After a series of community calls, Public Lab concluded members were most needed to scan open NOAA imagery and take screenshots of oil spills and flares. The conversations emerging from these initial calls on image sorting were out

¹⁹See Public Lab Store (2018) and How Public Lab is Funded (2018).

²⁰"How Public Lab is Funded."

²¹Willie, "Jakarta Flood Kite Mapping," February 5, 2014, <https://publiclab.org/notes/Willie/02-05-2014/jakarta-flood-kite-mapping>; Stevie, "Stormwater Workshop Two Report: Community Mapping," May 7, 2017, <https://publiclab.org/notes/stevie/05-02-2017/stormwater-workshop-two-report-community-mapping>.

²²The network visualization omits labels for the nodes in accordance with the aim of this project to ensure anonymity. The intended point—that the network is stabilized mainly by the contributions of a handful of gatekeepers and one clear cog—does not require them.

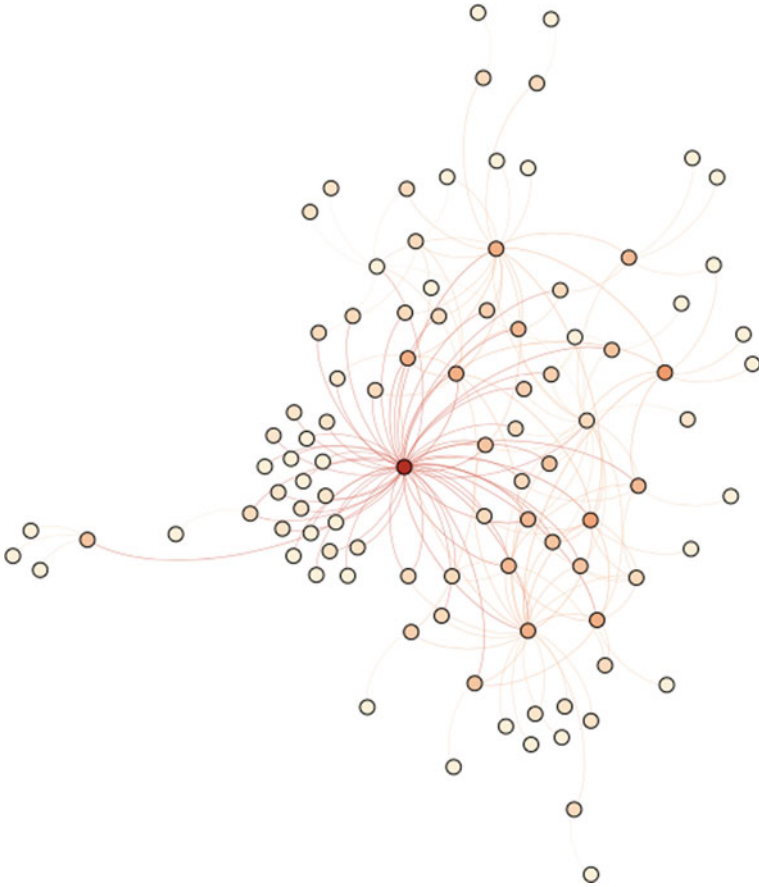


Fig. 10.6 A social network visualization of Public Lab research note contributions over a 6-month period. Node colors are deeper depending on the degree of the user’s contributions, which can also be seen in how well-connected a given node is to others in the visualization

of a need for cataloging footprints of buildings that are hazardous or have the potential to be hazardous that can be found close to flooded waterways. Landfills were also important to track for overrun and more naturally created landfill sites following the storm. A more recent but similar concern in Hurricane Florence’s aftermath is the potential for pollution from damaged industrial farms and coal ash sites.²³

Mapbox’s own effort in collaboration with OSM and DigitalGlobe, entitled “Hurricane Harvey: flooding, population, and known hazard map,” was also highlighted.²⁴ But the conversations leading up to the Harvey image sorting were broader in scope than Harvey in isolation. One concern voiced in its development, for instance, is

²³PBS Newshour (2018).

²⁴Hurricane Harvey: Flooding, Population, and Known Hazard Map (2018).

how so many facilities with hazardous emissions found themselves in such a vulnerable area, to begin with, and how emblematic it is of corporate power. One member noted that the clustering of these sites constitutes an “immense industrial no-man’s land” that is corporate-controlled and devastates wetlands areas toward further value creation. Indeed, from its Gulf Coast investments on, Public Lab work has long championed community-driven work monitoring wetlands loss.

There was also discussion about the need for outlets of citizen participation in these matters and the “right to know” about such hazards. Specifically, members discussed the weight of the Community Preparedness Right to Know Act in the area, given the government’s response (as one member noted) that not nearly enough documentation was provided to the public to move forward in any capacity under the act’s terms.

Public Lab was organizing specifically in response to a Coast Guard request for image links visualizing areas of concern in spreadsheet form with predefined columns and sources of information. Nonprofits associated with Public Lab flew over for images in Harvey’s aftermath but also needed help processing imagery. One member suggested Public Lab could easily make hackathons to support such requests, and Public Lab ran an image sorting workshop precisely to enlist people toward this work.

Public Lab previously experimented with solutions for such image processing tasks, most notably via the MapMill tool, a collaboration with HOT. It had two versions; while the first required users to input images on the backend, the second version manages web-based upload, but comes with various bugs. MapMill inspired Cartoscope, an image sorting citizen science mapping site utilized for disaster response but mostly for projects identifying algae bloom.²⁵ Additionally, Public Lab’s Cartography Collective, an established group of members who stitch often large archives of images provided by other community members to create maps, was voiced as a potential avenue toward completing a disaster imagery project.²⁶ Of these discussed options, Cartoscope proved arguably the most applicable; various spreadsheet entries identifying pollution sites give attribution to Cartoscope for assistance in locating what was found.

For the workshop, in searching through imagery for spills and flares, workshop attendees were told to copy URLs once they found areas of concern as a permalink containing latitude and longitude coordinates for what they spotted. The NOAA’s image set came from the National Geodetic Survey, highlighting FEMA and NWS designated sites between August and September 2017. The specific date of flight is available at the right-hand corner of the screen as one reviews imagery, which contrasts how the nature of HOT’s work often means the obfuscation of such helpful details, often to the annoyance of many in the HOT community.

This information was one of many vital elements recorded on the spreadsheet created for the initiative to conform with the Coast Guard guidelines. The spreadsheet had columns to note the name of the identified facility; the mapper who found it;

²⁵Cartoscope (2018).

²⁶Public Lab Cartography Collective (2018).

the link to access the imagery; latitude and longitude coordinates for the site; the date the image was taken; columns to indicate if the identified image is evidence of an oil sheen, site flooding, flare, or big stormwater discharge (to which users would respond by writing yes, no, or possibly under the appropriate column); and notes. For the latter, users often chose to convey particularities of the coloring or the flow of the water were they to fill in that field. The spreadsheet contains 183 entries by eight members, with identified sites including chemical plants, sewage plants, and oil refineries.

All in all, attendees identified and located 55 areas of potential concern in one evening alone. Efforts continued on a remote basis following the workshop, though members faced issues getting coordinates and using permalinks off NOAA's Harvey imagery to share what they were finding. While Public Lab members were able to reach out to the NOAA to patch the issue, there were still delays faced in the corrections going live and persisting in the use of the data thereafter. The typically closed-off nature of such data can thus still pose hindrances to such initiatives.

Though the workshop was a successful event, participants made suggestions for future workshops that coincide with many of the lessons learned from HOT tasks. The first recommendation was to designate an experienced participant with a role specifically in quality control to help out others so that they can be more efficient. A second recommendation was to work through examples of what to identify and how to use the involved interfaces at the start of the event. The third recommendation was to underscore that following the flow of a river can speak to the location of spills. These recommendations reflect a need, even on a smaller scale of participants than HOT, for clear instructions and modeling (as HOT attempted through its pilot internship program) of how to handle imagery in light of the unique challenges that reading aerial imagery can pose, especially for beginners. Instilling a sense of how to make aerial imagery legible means teaching mappers an appreciation for overlooked aspects of the environment that can illuminate elements of the built environment like buildings or aberrations of the natural environment like oil spills.

10.6 Conclusion

Though these largely self-organized disaster response campaigns are not explicitly how the US government imagined public mapping would be instantiated when it made GPS publicly available, the application matches the spirit of the inherent goals behind the move. It realizes a more efficient citizenry contributing to the work of government. This helps ensure the safety of those affected by the disaster at a time when government responses to disasters are under heavy critique within the media and the citizenry. There is a clear need for maps to be produced in this way in the face of a spike in natural disasters resulting from climate change and human activity. The rapid and unprecedented pace of these formations mandates actions that can prepare states and NGOs as best as possible once a formalized response is ready.

HOT is thoughtful about the future of this work and how it can facilitate a more seamless workflow. During a conversation with a HOT representative following my internship experience, I was told HOT's priority in the next decade should be moving from a management role to a supporting role, providing infrastructure when necessary and raising funds for communities to map on their own without warranting organizational requests. Instead of organizing "a hoard response" every time a base map is needed (which, in cases like Puerto Rico, meant corralling an estimated 4000 mappers in a matter of weeks), Missing Maps would have the capacity to build off preexisting data. This would include using an on-the-ground kite and UAV imaging uploaded to OpenAerialMap to generate crowd support from there—a potential avenue for the kinds of grassroots tactics to mapping Public Lab sponsors.

The scale, number of mappers, outlets by which members contribute, and means by which calls for contribution are established differs between the two communities discussed. The nature of the backend institutional relations behind the campaigns—pairing with governments, satellite initiatives and nonprofits—are similar. But they still manifest different approaches: one a matter of completion in standing base layer data that is open to use to assist in emergency routing, the latter problem-oriented and cultivated through significant community-wide dialogue.

The lessons this chapter offers for implementing crowdsourcing in aerial imagery assessment is that such work benefits from both modeling norms and particularities of the aerial imagery at hand for a particular campaign and underscoring the high degree of subjectivity this assessment entails. In discussing an experimental internship program and feedback from hackathons, this chapter demonstrates that planning in dedicated social interaction to highlight these particularities and judgment calls can aid such work significantly. Modeling actions within particular tasks, rather than relying solely upon broader task instructions, can ensure accuracy and efficiency. Without it, projects can either lag or face difficulties in quality control and impact. In addition, the mentorship of organizational leaders as a means of building in social relations can build members' motivation and confidence in providing such complex and valuable labor.

The prime research contribution of this chapter is to trace the social and technical means by which disaster mapping projects handle complex tasks in distributed and collaborative models. I provide findings that incorporate social networks, interface logistics, and contribution analytics within the case studies at hand. But this research also resonates broadly with various interests within research on crowdsourcing. These include the efficacy of different modes of peer feedback as well as the roles of curiosity and intrinsic motivation in increased contribution.

This chapter discusses both private modes of feedback through messages on the HOT interface as well as public modes of feedback through comments on tasks within the HOT Tasking Manager and disaster mapping events with Public Lab. While some see the former as "depersonalized," the latter often gets characterized positively for its potential in providing "social affirmation." But research on HOT identifies a need for a balance between the two, and indicates that receiving negative feedback via private messages does not impact the retention of mappers. It is the timing of the

feedback, the experience of the peer reviewer, and the adherence to certain norms within private feedback that research deems vital moving forward.²⁷

The internship experiences I recount exhibit these factors at work within the HOT community. Feedback can be slow through private messages; it may take less time for users to receive feedback if they request it specifically on the interface prior to saving their changes on the iD editor. The tenor of feedback when provided by other interns was fairly uniform, but comments from other validators ranged from curt to congratulatory.

Likewise, studies on crowdsourcing show that creating curiosity within task design can increase contributions and the quality of the resulting data.²⁸ User interest in exploring distant areas through this work may be a motivational factor. One can at the very least identify within HOT's calls for contribution moments in which HOT promotes this (perhaps problematically, given cartography's colonial history) as a persuasive tactic.

Contribution in the contexts this chapter examines is not merely a matter of curiosity, but equally one of perceived imperative.²⁹ Overall, intrinsic motivation factors like these remain relatively undertheorized within studies of crowdsourcing.³⁰ Thinking through how projects frame crowdsourced tasks in meaningful ways is especially worthy of attention. Research indicates meaningful framings increase the likelihood to participate in crowdsourced frameworks without compromising the quality of the work.³¹ This framing seems pivotal to how OSM contributions spike in the wake of crises, disasters, and epidemics.³²

When this work carries over into hackathons and mapathons, there is a need to measure the impact of social events on the platform, on the participant, and, if applicable, on the institution or requesting organization running the mapathon.³³ More exploratory and qualitative work similar to research on HOT that has identified a need for isolating the effect of norms and reviewers' level of experience in providing feedback on how these events model norms and behaviors can enrich understandings of how such platforms facilitate complex work.

Even at the level of remote individual contribution, research on macrotask crowdsourcing can benefit from the level of observational detail this chapter assumes. Perhaps the most important thread in these disaster mapping contexts is recognizing that reading aerial imagery means following various nuances to yield knowledge about the mapped environment. As this chapter shows, often overlooked but crucial natural elements such as tree cover, roads, or the color and flow of water imparts meaningful information that can help narrow down the locations of buildings or spills. Though

²⁷Dittus and Capra (2017, p. 40:18).

²⁸Law et al. (2016, pp. 4098–4110).

²⁹Anderson-Tarver (2015).

³⁰Zheng et al. (2011, p. 79).

³¹Chandler and Kapelner (2013, p. 123).

³²For research on OSM contributions in the event of earthquakes specifically, see Ahmouda et al. (2018, pp. 195–212).

³³Coetzee et al. (2018, pp. 41–42).

these are not the kinds of elements likely to garner media attention when disasters strike, they are becoming increasingly important when mappers organize to help those on the ground.

References

- About Us. *Youthmappers*. Retrieved September 3, 2018, from <https://www.youthmappers.org/about-us>.
- Ahmouda, A., Hochmair, H. H., & Cvetojevic, S. (2018). Analyzing the effect of earthquakes on OpenStreetMap contribution patterns and tweeting activities. *Geo-spatial Information Science*, 21(3), 195–212. <https://doi.org/10.1080/10095020.2018.1498666>.
- Anderson-Tarver, C. (2015). Crisis mapping the 2010 earthquake in OpenStreetMap Haiti. *Geography Graduate Theses & Dissertations*.
- Arroy, B. Biota survey for Devils Lake, ND. *George W. Bush White House Archives*. Retrieved December 23, 2016, from https://georgewbush-whitehouse.archives.gov/ceq/biota_survey_200507.html.
- Cartoscope. Retrieved July 16, 2018, from <http://cartosco.pe/#/home>.
- Chandler, D., & Kapelner, A. (2013). Breaking monotony with meaning: Motivation in crowdsourcing markets. *Journal of Economic Behavior and Organization*, 90, 123–133.
- Charles G. Groat hosts ask the White House. *George W. Bush White House Archives*. Retrieved October 11, 2004, from <https://georgewbush-whitehouse.archives.gov/ask/text/20041011.html>.
- Coetzee, S., Minghini, M., Solis, P., Rautenbach, V., Green, C. (2018). Towards understanding the impact of mapathons—Reflecting on Youthmappers experiences. In *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XLII-4/W8*, FOSS4G 2018.
- Creative Commons Attribution 4.0 International Public License. *Creative Commons*. Retrieved August 13, 2018, from <https://creativecommons.org/licenses/by/4.0/legalcode>.
- Dittus, M., Capra, L. (2017). Private peer feedback as engagement driver in humanitarian mapping. In *Proceedings of ACM Human Computer Interaction*, 1, 40:1–40:18.
- Dolan, T. E. (2010). Revisiting adhocracy: From rhetorical revisionism to smart mobs. *Journal of Futures Studies*, 15(2), 33–50.
- Grassroots Mapping Forum, Issue 1. <https://i.publiclab.org/system/images/photos/000/005/051/original/gmf-1-prepress-small.pdf>.
- Hass, D., Ansel, J., Gu, L., Marcus, A. (2015). Argonaut: Macrotask crowdsourcing for complex data processing. In *Proceedings of the 41st International Conference on Very Large Data Bases*, 8(12), 1643–1644. <https://dl.acm.org/citation.cfm?doid=2824032.2824062>.
- How Public Lab Is Funded. Retrieved April 18, 2018, from <https://publiclab.org/wiki/how-we-are-funded>.
- Hurricane Harvey: Flooding, Population, and Known Hazard Map. *Mapbox*. Retrieved July 16, 2018, from <https://www.mapbox.com/bites/00368/#14.27/29.7032/-95.3549>.
- Jemielniak, D. (2014). *Common knowledge? An ethnography of Wikipedia*. Stanford: Stanford University Press.
- Kamensky, J. *A brief history of Vice President Al Gore's National Partnership for Reinventing Government during the administration of President Bill Clinton 1993–2001*. Retrieved January 12, 2001, from <http://govinfo.library.unt.edu/npr/whoware/historyofnpr.html>.
- Law, E., Yin, M., Goh, J., Chen, K., Terry, M., Gajos, K. Z. (2016). Curiosity killed the cat, but makes crowdwork better. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 4098–4110).
- Missing Maps. Retrieved June 11, 2018, from <http://www.missingmaps.org>.

- National Day of Civic Hacking. *Code for America*. Retrieved September 13, 2018, from <https://www.codeforamerica.org/national-day-of-civic-hacking>.
- OSM Analytics. Retrieved June 25, 2018, from <http://osm-analytics.org/#/>.
- Ouellette, L., Hay, J. (2008). *Better living through reality TV: Television and post-welfare citizenship*. Malden: Wiley.
- Partnerships. *Humanitarian OpenStreetMap Team*. Retrieved October 30, 2017, from <https://www.HOT.org/partnerships>.
- PBS Newshour. *North Carolina Florence flood disaster is growing*. YouTube, 3:57. Retrieved September 17, 2018, from <https://www.youtube.com/watch?v=GcjmzHoiGU>.
- President Discusses War on Terror at Naval Academy Commencement. *George W. Bush White House Archives*. Retrieved May 27, 2005, from <https://georgewbush-whitehouse.archives.gov/news/releases/2005/05/20050527.html>.
- Press Briefing by FEMA Director David Paulison. *George W. Bush White House Archives*. Retrieved July 31, 2006, from <https://georgewbush-whitehouse.archives.gov/news/releases/2006/07/20060731-4.html>.
- Public Lab Cartography Collective. Retrieved July 16, 2018, from <https://publiclab.org/wiki/public-lab-cartography-collective>.
- Public Lab Store. Retrieved April 18, 2018, from <https://publiclab.org/wiki/how-we-are-funded>.
- PUBLIC LAW 107-347—DEC. 17 2002. Retrieved April 26, 2017, from <https://www.gpo.gov/fdsys/pkg/PLAW-107publ347/pdf/PLAW-107publ347.pdf>.
- Remarks by President George Bush, Prime Minister of Ireland Bertie Ahern, and President of the European Commission Romano Prodi in Press Availability. *George W. Bush White House Archives*. Retrieved June 26, 2004, from <https://georgewbush-whitehouse.archives.gov/news/releases/2004/06/20040626-16.html>.
- Radiant Earth*. Retrieved May 24, 2018, from <https://www.radiant.earth>.
- Richard Armitage hosts ask the White House. *George W. Bush White House Archives*. Retrieved July 2, 2004, from <https://georgewbush-whitehouse.archives.gov/ask/20040702.html>.
- Scoles, S. (2017, August 28). High-res satellites want to track human activity from space. *Wired*. https://www.wired.com/story/high-res-satellites-want-to-track-human-activity-from-space?_lsrc=71a78e96-c497-4126-93ef-37274fa9a3fb.
- Segal, C. Volunteers are helping Puerto Rico from home, with a map that anyone can edit. *PBS Newshour*. Retrieved October 1, 2017, from <http://www.pbs.org/newshour/rundown/volunteers-helping-puerto-rico-home-map-anyone-can-edit/#.WdJoSMearCI.twitter>.
- The President's New Freedom Initiative: The 2007 Progress Report. *George W. Bush White House Archives*. Retrieved April 26, 2017, from <https://georgewbush-whitehouse.archives.gov/infocus/newfreedom/newfreedom-report-2007-3.html>.
- Urban Innovations: Crowdsourcing Non-Camp Refugee Data. *Humanitarian OpenStreetMap team*. Retrieved October 31, 2017, from https://www.HOT.org/projects/urban_innovations_crowdsourcing_non_camp_refugee_data.
- U.S.—Canada Smart Border/30 Point Action Plan Update. *George W. Bush White House Archives*. Retrieved December 6, 2002, from <https://georgewbush-whitehouse.archives.gov/news/releases/2002/12/20021206-1.html>.
- Warren, J. (2010). *Grassroots mapping: Tools for participatory and activist cartography*. Master's thesis, MIT.
- Wilkinson Bay, Louisiana. July 22, 2010. CC-BY-SA 2010.
- Yin, A. (2017, October 2). A mapathon to pinpoint areas hardest hit in Puerto Rico. *The New York Times*. <https://www.nytimes.com/2017/10/02/nyregion/maps-puerto-rico-hurricane-maria.html>.
- Zheng, H., Li, D., & Hou, W. (2011). Task design, motivation, and participation in crowdsourcing contests. *International Journal of Electronic Commerce*, 15(4), 57–88.