CHI 2018 Hackathon Workshop Submission: Initial Findings from Hackathon Trace Data

Erin Hoffman
3rd-Year PhD Student
University of Washington
Seattle, WA 98195, USA
erinrhof@uw.edu

Paste the appropriate copyright/license statement here. ACM now supports three different publication options:

- ACM copyright: ACM holds the copyright on the work. This is the historical approach.
- License: The author(s) retain copyright, but ACM receives an exclusive publication license.
- Open Access: The author(s) wish to pay for the work to be open access. The additional fee must be paid to ACM.

This text field is large enough to hold the appropriate release statement assuming it is single-spaced in Verdana 7 point font. Please do not change the size of this text box.
Each submission will be assigned a unique DOI string to be included here.

Abstract

Competitive overnight coding and prototyping events known as "hackathons" represent a large and growing phenomenon, with tens of thousands of participants and millions of dollars spent each year. The social computing community has made progress in understanding certain hackathons through ethnographic and research-through-design methods, but has not yet answered a number of important questions about hackathons as a whole. This paper analyzes hackathon-related trace data from the websites Devpost.com and GitHub.com to address some of these questions about the geographical distribution of hackathons, the distribution of projects across hackathons, long-term hackathon outcomes, the content of hackathon projects, and hackathon participant networks. This analysis has generated four data visualizations and seven insights from those data visualizations. These visualizations and insights form the first step in a larger mixed-methods investigation.

Author Keywords

Hackathons; Trace Data; Mixed Methods

Introduction

I have been a hackathon attendee and organizer since 2012 and am in the process of conducting a mixed-methods investigation of hackathon trace data. I am thus both a hackathon practitioner and a hackathon researcher. I hope to contribute to the 2018 CHI "Hacking and Making at Time-Bounded Events" workshop in the contexts of *design variations, short-term and long-term outcomes, practical support for hackathon organizers,* and *theoretical space of hackathons*.

Experience as a Hackathon Attendee and Organizer

My first hackathon was a Windows Phone hackathon in 2012, which was hosted by Microsoft representatives and took place in Michigan State University's Computer Science and Engineering department conference room. It was a small event with Jimmy John's catering where I first heard someone use the term "hack" to mean "a clever but brittle quick fix" rather than "a data breach". My teammate and I were the only women in attendance, and we won the people's choice award (and two Windows Phones and an Xbox) for our prototype (which tried to improve Bing Maps on campus). I left feeling like I hadn't wasted my Saturday, but didn't anticipate going to many more hackathons.

When a close friend persuaded me to go to HackIllinois in 2014, I changed my mind. Instead of 30 people and a handful of projects in a conference room, HackIllinois had hundreds of attendees and dozens of projects spread out over multiple campus buildings at the University of Illinois Urbana-Champaign. We were alternating between hours of coding and debugging and sleeping on the hard floor of a brightly-lit classroom, but the warm community and collective effervescence [1] were delightful. I also felt that I had "leveled up" as an engineer by learning how to use tools we had never used in class (like REST APIs and GitHub), and by networking (and playing cards) with employees of elite tech companies. This hackathon experience made me realize that I really enjoyed this type of large hackathon, which is also known as a "collegiate hackathon" and is often a member event of an organization called Major League Hacking.

Since HackIllinois in 2014, I have competed and mentored at dozens of collegiate and other hackathons across the United States, and in the past year also mentored at a hackathon in Abu Dhabi, UAE. I also helped to found SpartaHack, Michigan State University's collegiate hackathon, which welcomed 300 participants in 2015 and successfully completed its 600-person fourth edition in 2018. More recently I helped organize DubHacks, the University of Washington's collegiate hackathon. As of February 2018 I have experienced thousands of hours as a hackathon attendee, volunteer, mentor, and organizer.

Research Interest in Hackathons

When I began my PhD, I didn't intend to study hackathons. I planned to build tools for moderators on sites like Wikipedia, Reddit, and Facebook – tools that would use natural language processing to provide a quantified basis for removing harmful content. After spending more than a year investigating possibilities, I concluded that accuracy standards in computational linguistics are too different from accuracy standards in online content moderation for this kind of tool to be viable with today's computational linguistics tools [2]. Maybe I will return to this project in a few years when

hackathons are better understood by the research community!

I first turned to hackathons as a research topic when I took a qualitative methods class. I conducted a partial Grounded Theory [3] participant observation study of my fellow hackathon organizers at DubHacks, and found tensions and contradictions around how organizers conceptualized the purpose of hackathons [4].

While performing the literature review for that project, I found that relatively few papers had been written about hackathons thus far, and that almost all of them have been qualitative or mixed-methods studies with small numbers of participants [e.g. 5, 6, 7] and/or research-through-design studies where the researchers participated in the organization of a hackathon geared towards some particular design outcome [e.g. 8, 9, 10]. Most of these studies investigated only one hackathon, and all (that I could find) investigated fewer than ten hackathons. While these kinds of studies are able to answer interesting questions, their narrow scoping prevents them from painting a broader picture of the hackathon phenomenon. I want to be able to answer larger questions about hackathons like:

- What do people build at hackathons?
- Why do people build what they build at hackathons?
- Are hackathon projects discarded after the hackathon ends, or do they become long-term projects?
- Who participates in hackathons? How often do they participate more than once, and what factors predict this?
- Are hackathon participants likely to learn new skills at hackathons that they will apply in future projects?
- How do prizes impact all of the above?

Many or most of these questions would be practically impossible to answer in a generalizable way with the research methods that have been applied to hackathons thus far. Almost all of the previous work has relied heavily on interview data, which provides rich insight but cannot reasonably scale to dozens or hundreds or thousands of hackathons.

Fortunately, there is another source of data that researchers have not yet utilized: online trace data. As a hackathon participant, I learned how participants use the websites Devpost.com and GitHub.com to showcase and store their work. As a researcher, I have begun to utilize the traces left on these websites as a data source to help address these big questions.

Methods

Under the instruction of Kate Starbird, I am working on a mixed-methods research project with a similar methodology to her paper titled "(How) Will the Revolution be Retweeted?" [11]. This methodological process involves multiple iterations of quantitative data/network analysis and qualitative grounded theory analysis in order to construct a broad and deep understanding of a set of online trace data. This process is currently a work in progress at the initial quantitative stage, and I will present my current findings in this paper.

Data Collection

In order to begin to address my questions about hackathons more broadly, I collected data from the Major League Hacking (MLH) Fall 2017 season.[1] Major League Hacking is an organization that supports hundreds of college and high school hackathons, and I used its event listing as a starting point. Eventually, I plan to collect data from a significantly larger number of hackathons.

Based on the listings on the MLH website, I identified 79 hackathons that took place in North America between August and December 2017. For each of the 79 hackathons, I collected (or attempted to collect) the name, location, link to the Devpost.com submissions page, and the timestamp of when winners were announced.

Of these 79 hackathons, 73 had submissions on Devpost.com. Devpost is a software portfolio website that also allows hackathon participants to submit summaries of their projects for hackathon judging. Each project submission on Devpost has information about the project, such as a name, tagline, and description, and some Devpost project submissions also have links to GitHub.com repositories. GitHub is an online source control tool that allows users to post their code as well as a history of changes to their code. From the starting point of 73 MLH hackathons, I found 4,637 project submissions on Devpost. Of those 4,637 project submissions, 2,742 had links to GitHub, although only 2,506 of these links led to public GitHub repositories.

[1] https://mlh.io/seasons/na-2018/events (Fall 2017 events are part of the "2018 season")

For each of the 4,637 project submissions, I collected (or attempted to collect) the name, tagline, description, "built with" tags, Devpost link, GitHub link, hackathon name, and links to up to 6 author profiles on Devpost (plus a count of total authors).

For each of the 2,506 public GitHub repositories, I collected the number of commits (a term describing a set of changes made to a repository) both before winners were announced for the hackathon and after winners were announced for the hackathon.

Data collection was completed in January 2018, so some data may be stale. For example, some Devpost users might have edited their submission descriptions, or some GitHub users might have made additional commits.

Analysis

Following my mixed-methods trace data methodology, I used data visualization tools to get a broad overview of the data I collected. This is the first step in the process, and my next step will be to begin a grounded analysis of the content of the project submission descriptions on Devpost and the GitHub commits.

Results

In order to get a high-level view of my data, I created four different data visualizations. The first is a map view with each hackathon represented by a circle, showing the geographical distribution of these events. The second is a bar graph with each hackathon along the x-axis and the number of projects and number of long-term projects represented on the y-axis. The third is a word cloud showing the most common tags listed in the "built-with" field. The fourth is a network graph, showing collaboration links between Devpost users.

Insight 1: The majority of hackathons and hackathon project submissions in this dataset are in the eastern half of the US, not in or near Silicon Valley

Insight 2: The largest hackathons by number of submissions are clustered at some prestigious universities with strong engineering programs:

- University of Washington (DubHacks)
- University of California, Berkeley (Cal Hacks 4.0)
- University of California, San Diego (SD Hacks)
- University of Michigan (MHacks X)
- Georgia Tech (HackGT)
- University of Waterloo (Hack the North)
- University of Pennsylvania (PennApps)
- Princeton University (HackPrinceton)
- Yale University (YHack)
- Harvard University (HackHarvard)

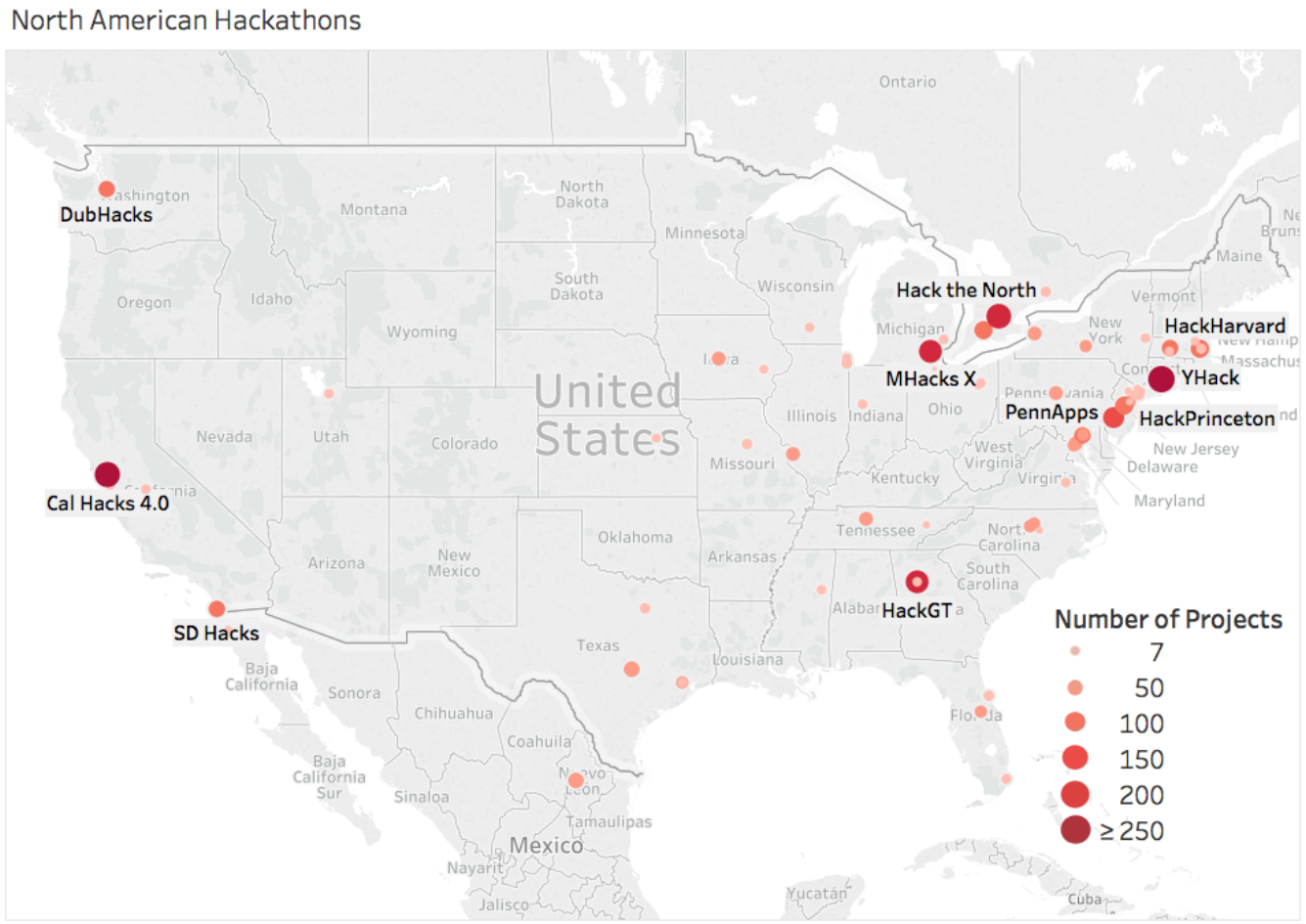

Figure 1: This visualization shows the distribution of hackathons across North America (USA, Canada, and Mexico) for the Fall 2017 season. Each circle representing a hackathon is larger and darker depending on the number of projects submitted at that hackathon. This visualization was created using Tableau.

Insight 3: The majority of hackathons have fewer than 100 submissions. The minimum is 7, maximum is 271, mean is 63.52, and median is 44.00.

Insight 4: A fairly small proportion of projects have any commits on GitHub from after their respective hackathons ended.

- Overall, 11.5% of projects had GitHub commits after their hackathons ended
- Out of the 2,506 projects with public GitHub repository links, 21.2% had GitHub commits after their hackathons ended
- Purdue University's hackathon (BoilerMake) had the highest percentage of projects continued (32.4%)
- 8 hackathons out of 73 had 0 projects continued
- Earlier hackathons (further left on the graph) seem to be slightly more likely to have projects be continued

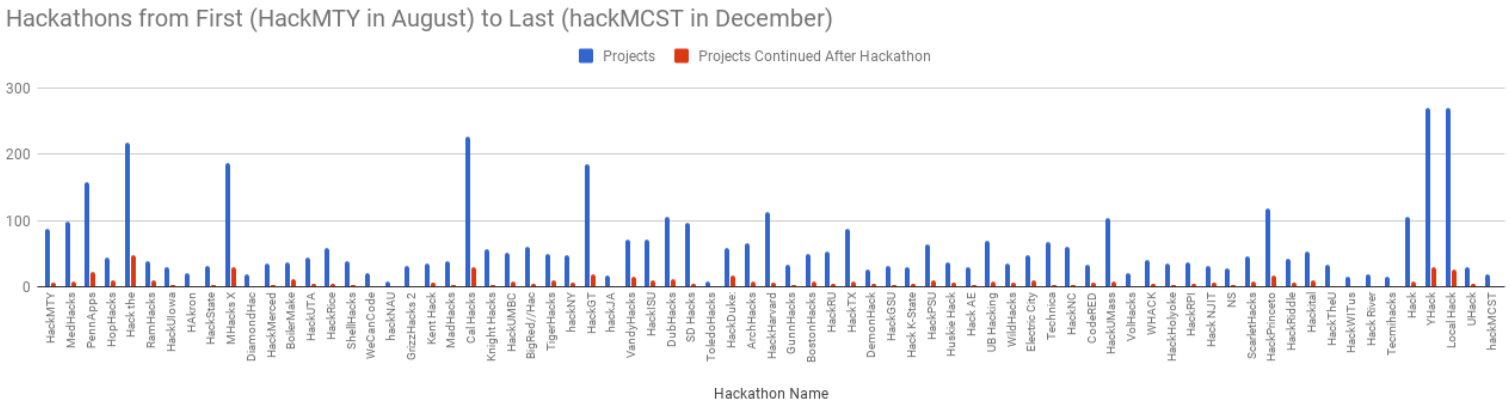

Figure 2: This visualization shows the distribution of projects and continued projects for each hackathon. Hackathons are arranged horizontally in chronological order, starting with HackMTY in August and ending with hackMCST in December. Blue bars represent the number of projects submitted on Devpost, and red bars represent the number of projects with at least one GitHub commit (set of code edits) after the hackathon's winners were announced. This is likely to be an underestimate of how many projects have actually been continued, because it only includes projects with publicly-available GitHub repositories. This visualization was created using Google Sheets.

Insight 5: The "built with" tags are clustered around 3 tags (N.B.: these are percentages of the projects whose authors filled in the "built with" field and did not leave it blank):

- "javascript" was used in 1,580 out of 4,480 projects with tags (35.3%)
- "html" was used in 1,474 out of 4,480 projects with tags (32.9%)
- "python" was used in 1,458 out of 4,480 projects with tags (32.5%)

Figure 3: This visualization shows the distribution of tags used in the Devpost projects' "built with" field. Larger font size indicates the tag was used more times. (N.B.: only the most popular tags are displayed out of 520 distinct tags.) This visualization was created using WordItOut.com.

Insight 6: The large clusters near the center represent highly active hackathon participants who frequently form teams with others. The largest, darkest circle represents a user who collaborated with 19 different people over the course of 7 different hackathons.

Insight 7: Most people had collaborated with 2-3 other people (N.B.: this graph only includes people who worked with at least one person)

- 84.3% worked with 2 or more people
- 57.6% worked with 3 or more people
- 15.4% worked with 4 or more people
- 5.5% worked with 5 or more people
- 2.8% worked with 6 or more people

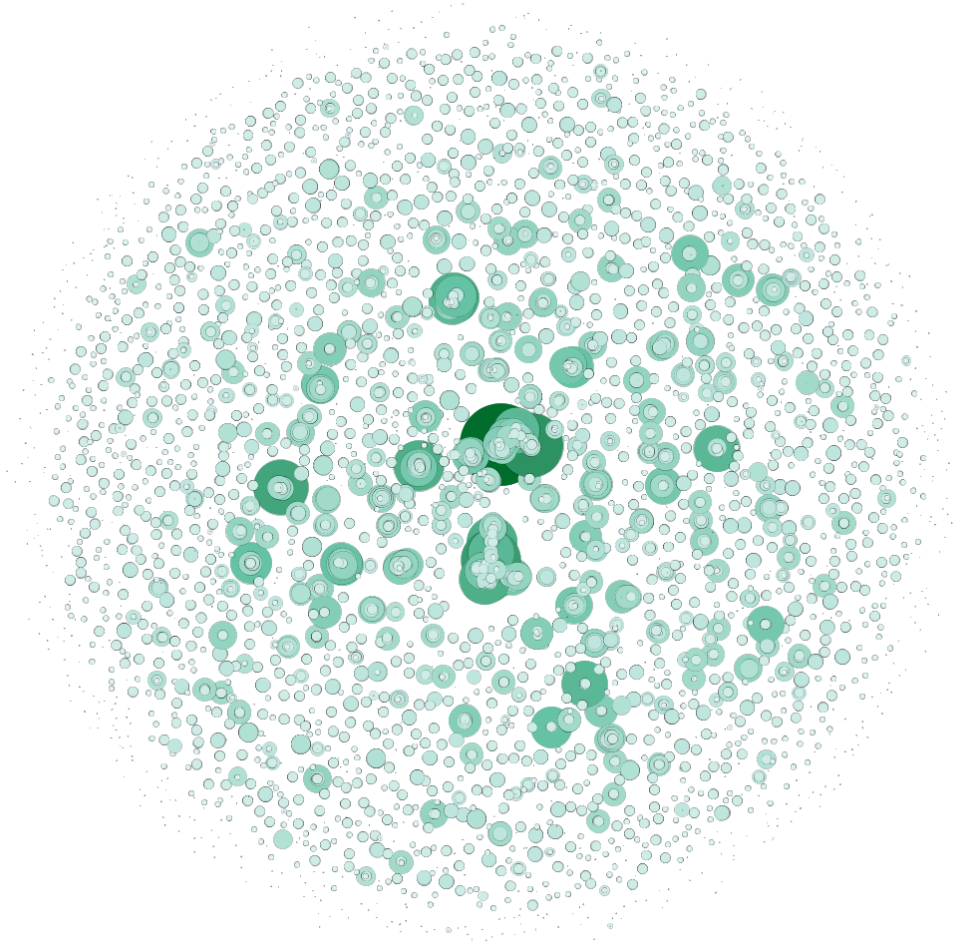

Figure 4: This visualization shows collaborations between participants at hackathons based on Devpost submissions. Each circle represents a participant, and the circles are larger and darker when the participants have collaborated with more people. Circles are spatially located nearer to the circles of the people they have collaborated with and further from the circles of the people they have not collaborated with. This visualization was created using Gephi.

Conclusion

Overall, the four visualizations have led me to seven insights. (I use the term "insight" rather than "finding" or "conclusion" because this is only a preliminary step.)

First, MLH hackathons in Fall 2017 were geographically clustered around the eastern half of the United States, with a particularly dense cluster going from Washington, DC to Cambridge, MA.

Second, the largest of these events (by number of project submissions) took place at prestigious universities with strong engineering programs. Despite this, some prestigious universities with strong engineering programs had relatively small hackathons (e.g. Carnegie Mellon University, Northwestern University).

Third, these hackathons each had between 7 and 271 project submissions, with a median of 44 project submissions. This was smaller than my original guess, which was a median of 60 or 70 project submissions.

Fourth, a fairly small percentage of hackathon projects appear to be long-term projects. When "long-term" is defined as "having at least one commit on GitHub after hackathon winners are announced", only 21.2% of projects with publicly-available GitHub commit data were long-term.

Fifth, out of 520 distinct "built with" tags, "javascript", "html", and "python" were the most popular. Each was present in about 1/3 of all Devpost submissions where the "built with" field was filled out.

Sixth, a few participants attended many hackathons and formed large numbers of connections with different collaborators, the highest being 19 collaborators. The participant with 19 different collaborators attended 7 of the 73 hackathons in the dataset.

Seventh, and finally, most participants who collaborated with at least one other person collaborated with at least 3 other people (57.6%). A very small proportion of participants collaborated with at least 6 other people (2.8%).

The next step of this project will be a qualitative grounded theory analysis of the actual content behind these numbers. Following the advice of Howard [12], I will use insights from these quantitative visualizations to guide me in sampling the qualitative data. For example, I might investigate the kinds of projects created at BoilerMake in order to better understand why that particular hackathon had the highest rate of long-term projects. I also might look into the differences between the small minority of people who have had more than five collaborators and the large majority of people who have had three or fewer. (In particular, I am interested in finding out whether this is tied to geography – the people with the highest numbers of collaborators seem to live on the East Coast where the geographic density of hackathons is high.) Finally, I might investigate what kinds of projects include the most popular "built with" tags ("javascript", "html" and "python") and what kinds of projects include more-obscure tags like "photoshop" or "love".

At this CHI workshop, I hope to share these initial findings with the group and to get your feedback on what to investigate next.

References

1. Durkheim, Emile and Joseph Ward Swain. *The Elementary Forms of the Religious Life*. 1912.
2. Hoffman, Erin R., David W. McDonald, and Mark Zachry. "Evaluating a Computational Approach to Labeling Politeness: Challenges for the Application of Machine Classification to Social Computing Data." In *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW, Article 52 (December 2017).
3. Charmaz, Kathy. *Constructing Grounded Theory*. Sage, 2014.
4. Hoffman, Erin. "Collegiate Hackathons as Liminal Spaces." Paper presented at Hacking and Making at Time-Bounded Events workshop at CSCW 2017.
5. Trainer, Erik H., Arun Kalyanasundaram, Chalalai Chaihirunkarn, and James D. Herbsleb. "How to hackathon: Socio-technical tradeoffs in brief, intensive collocation." In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, pp. 1118-1130. ACM, 2016.
6. Porter, Emily, Chris Bopp, Elizabeth Gerber, and Amy Voida. "Reappropriating Hackathons: The Production Work of the CHI4Good Day of Service." In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pp. 810-814. ACM, 2017.
7. Warner, Jeremy and Philip J. Guo. "Hack .edu: Examining How College Hackathons Are Perceived by Student Attendees and Non-Attendees." In *Proceedings of the 2017 ACM Conference on International Computing Education Research*. ACM, 2017.
8. Birbeck, Nataly, Shaun Lawson, Kellie Morrissey, Tim Rapley, and Patrick Olivier. "Self Harmony: Rethinking Hackathons to Design and Critique Digital Technologies for Those Affected by Self-Harm." In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pp. 146-157. ACM, 2017.
9. Clark, Russ, Matt Sanders, Brian Davidson, Siva Jayaraman, and Carl DiSalvo. "The Convergence Innovation Competition: Helping Students Create Innovative Products and Experiences via Technical and Business Mentorship." In *International Conference on Human-Computer Interaction*, pp. 144-153. Springer International Publishing, 2015.
10. Lamela, Zapico, Jorge Luis, Daniel Pargman, Hannes Ebner, and Elina Eriksson. "Hacking sustainability: Broadening participation through green hackathons." In *Fourth International Symposium on End-User Development. June 10-13, 2013, IT University of Copenhagen, Denmark*. 2013.
11. Starbird, Kate and Leysia Palen. "(How) Will the Revolution be Retweeted?: Information Diffusion and the 2011 Egyptian Uprising." In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*. ACM, 2012.
12. Howard, Philip N. "Network Ethnography and the Hypermedia Organization: New Media, New Organizations, New Methods." In *New Media & Society*, 4(4) pp. 550-574.