THIS DOCUMENT IS A PREPRINT.

THE PUBLISHED VERSION IS:

FOOTE, J., & HWANG, S. (2023). ONLINE COMMUNITIES AND BIG DATA. IN T. REIMER, E. S. PARK, & J. A. BONITO (EDS.), GROUP COMMUNICATION: AN ADVANCED INTRODUCTION. ROUTLEDGE.

Online Communities and Big Data

Jeremy Foote
Purdue University

Sohyeon Hwang
Northwestern University

OBJECTIVES

1. Understand the key features of online communities.
2. Understand how group processes are influenced by the features of online communities.
3. Understand how digital trace data can be used (and misused) in studying online communities.

INTRODUCTION

In January of 2021, a number of members of the Reddit community 'r/wallstreetbets' argued that stock in the retail video game store GameStop was underpriced. Perhaps more importantly, they pointed out that wealthy hedge funds had "shorted" the stock. When shorting a stock, a hedge fund borrows shares of the stock from a broker and sells them immediately, hoping to buy the stock later at a lower price so they can return the shares to the broker and pocket the difference. Members of r/wallstreetbets argued that if the community banded together to buy GameStop shares then they could drive the stock price up. This would force the hedge funds into a "short squeeze" where they would

have to buy the stock back at the new higher price, driving the price up even further. Thus, community members could both make money *and* stick it to the wealthy hedge funds.

Amazingly, their plan mostly worked. As r/wallstreetbets members bought GameStop stock the price rose sharply and when deadlines for the hedge funds to return the borrowed stock grew closer, r/wallstreetbets members convinced each other to avoid selling the stock (and cashing in on their gains). In the end, a number of hedge funds lost a lot of money and a number of r/wallstreetbets members made a lot of money (Phillips et al., 2021)—although like many bubbles, the price eventually fell and other community members lost money. The chaos created by the community's actions led to emergency stock freezes, Congressional hearings, and class-action lawsuits.

So how did a group of strangers, communicating through a simple web forum, manage to coordinate and motivate their members to the point that they were willing to take huge financial risks? The case of r/wallstreetbets is in many ways an anomaly, but it serves as a clear example of the impact and potential of online communities. Countless other online communities serve as important gathering places for people to socialize, seek and share information, and collaborate on shared projects. Millions of people spend millions of hours writing code, editing articles, and engaging in public conversations. In this chapter, we will talk about what online communities are, some key features that influence how they operate as groups, and how the data from online platforms is enabling exciting new group-based research.

What are online communities?

We define online communities as virtual spaces where people freely and voluntarily convene around a shared interest. A virtual "space" is a communal online interface that allows interaction between group members, such as forums, wikis, or GitHub projects (see Figure 1). In online communities, the shared space is often (but not always) public and all group members experience it in basically the same way. Communities vary in their particular social structure and size, as well as in what interface they use to communicate.

Contemporary online communities take many different forms, which is part of what makes them so interesting. For example, question and answer sites like Yahoo! Answers or StackOverflow, teams of online gamers, online learning platforms like Scratch, or even the comment sections of news articles, gaming streams, or blogs could all be considered online communities. Despite

Figure 1: An example of an online community. On the subreddit "r/science", participants share and discuss peer-reviewed academic articles.

this diversity, not all online interactions happen in online communities. For example, two other prominent types of online interactions include: virtual teams, which are distributed work teams who coordinate and communicate using online tools, but are obligated to do so as part of their employment; and social media (like Instagram, Snapchat, or TikTok), where individuals experience a personalized "stream" of content.

Attributes of Online Communities

One way of thinking about how groups work is the **Input-Process-Output model** (Ilgen et al., 2005). According to this model, groups have a set of inputs, including their skills, attributes, resources, and raw materials. Through communication and other group processes, they transform these inputs into outputs. We use this model as a way of organizing some of the key features of online communities. In *inputs*, we discuss why and how people participate in online communities; in *processes*, we discuss the social and technical processes underpinning interactions and why those matter; and in *outputs*, we discuss

the typical outcomes and consequences of online communities.

Input

The primary inputs of a group are its members' time, skills, knowledge, and contributions, such as posts, comments, or messages. Online communities differ wildly from many other types of groups when it comes to how difficult it is to participate, how visible membership is to others, who is allowed to participate, and what motivations group members have.

Degrees of engagement Often, joining an online community is as easy as clicking a button. The low barriers to entry and exit in online communities necessitate a broad definition of participation and membership. In face-to-face groups, for example, it is very clear who is present and participating in a group. In online communities, on the other hand, membership can be basically invisible. Indeed, often the vast majority of participants of online communities are **lurkers** who consume content without posting in the community (Nonnecke & Preece, 2000). Even among those who do contribute content, there are stark differences in degrees of engagement. While most participants contribute very little, a few participants may spend hours each day contributing content to a community. This pattern of participation inequality is surprisingly consistent across online communities (see Figure 2). It is tempting to see lurkers and low-effort contributors as free-riders who benefit from the efforts of others (c.f., Olson, 1965). In many contexts, those free-riding on group efforts may be subject to social or formal sanctions, but the relative invisibility of lurkers and the low barriers to exit make some types of sanctions both more difficult and less effective in online communities (Gibbs et al., 2021). On the other hand, lurkers may be seen as the "audience" for the content produced and a larger audience can encourage greater participation by active participants (Zhang & Zhu, 2011).

Low costs of joining and leaving—combined with the fact that communities often exist on platforms—enable people to participate in *multiple* communities much more easily than is possible in face-to-face groups. This mode of engagement is also substantively different from participation in virtual teams, where team membership is typically more restricted and less flexible. One implication of this is that just as it is difficult to draw a clear line around who is a member of an online community it is also difficult to draw clear lines between different communities—the boundaries between them can be indistinct as conversations and members move between communities.

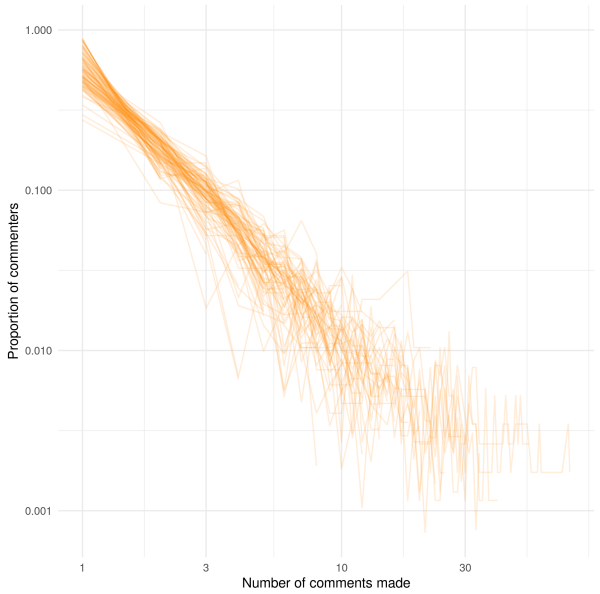

Figure 2: This figure from Foote (2022) shows the number of comments per person for 100 randomly selected subreddits active in January 2017, with the most prolific 5% of users removed. The x-axis shows the number of comments, and the y-axis is the proportion of commenters making that number of comments. In every single case, the vast majority of contributors make only a few contributions. Both axes are logged, so the distribution is even more unequal than it looks.

Anonymity and pseudonymity The virtual nature of online spaces makes it possible for participants to be **pseudonymous** (or even entirely anonymous) in ways that are usually impossible in other kinds of groups. While we might think that this would simply make relationships more difficult and norms more difficult to enforce, the influence of anonymity is complicated. In some cases, anonymity does have a disinhibition effect, leading to anti-social behavior (Suler, 2004). However, anonymity can also help people to control what aspects of their identity they reveal, and to whom, allowing them to participate in groups that they don't yet want to attach to their offline identities (e.g., LGBTQ+ groups providing support for people who haven't yet come out) (Ammari et al., 2019). We discuss additional impacts of anonymity on group processes in the *process* section below.

Motivations for participation One argument behind why online communities succeed is that they are able to harness the contributions of people who contribute for very different reasons (Benkler, 2002). *Uses and gratifications theory* proposes that people intentionally seek out multiple types of media in

order to satisfy different needs (Ruggiero, 2000). Researchers have found that participants seek out different online communities to fit their diverse needs (TeBlunthuis et al., 2022). Broadly speaking, researchers have identified many different motivations for participating in online communities, including connecting with others, information-seeking and -sharing, entertainment, and feeling a sense of belonging (Lampe et al., 2010; Ren et al., 2012). It's important to note one motivation that is missing—while most production-oriented offline groups (like work teams) use financial compensation as a primary motivation, even in online communities focused on production financial compensation is rare.

Process

Most groups are trying to do something. This might be as simple as entertaining other group members, or as complicated as building software or designing rockets. In order to meet those goals, groups need to coordinate who does what, make group decisions, socialize newcomers, etc. In offline groups, much of this work happens through synchronous meetings and interactions. In online communities, processes are often remote and asynchronous, mediated by user interfaces, algorithms, and platforms, and participated in by a rotating host of volunteers.

Communication tools Much of the communication in an online community happens directly, through posts, comments, and talk pages. Treem and Leonardi (2013) argue that online communication tools have four attributes that have the potential to dramatically influence how groups organize: visibility, persistence, editability, and association. Visibility refers to the fact that conversations and actions are often accessible and searchable by others, providing new ways to share knowledge about who knows what, also called "transactive memory" (Wegner, 1987). Persistence refers to the way that communications can remain available long after they were created, making them valuable shared public repositories that can be retrieved and built upon over time. Editability refers to both the ability of group members to think carefully and edit messages before making them visible as well as the ability in most group software to edit communications after posting. This can help group members to control how they present themselves and how they are perceived by others. Finally, association refers to the visibility of connections between different group members or between group members and content.

Treem and Leonardi (2013) argue that among other influences, association can make social relationships more likely and increase social capital.

Shared, persistent content provides another unique communicative opportunity for online communities called **stigmergic communication**. Stigmergy is the idea that communication can happen through an artifact itself (Heylighen, 2016). This is most obvious for open source software or wiki communities, where the community has a clear shared artifact they are working on together. For example, a wiki community member might make a link to a non-existing page to signal that someone should to create it. Stigmergy-like interactions occur even in conversation-based communities, where interactions are often mediated by algorithms. For example, many conversation-based communities like Reddit rely on non-linguistic communication in the form of reactions and likes, upvotes/downvotes, or sharing in order to determine what content to prioritize or to hide. At times the primary interaction of online communities may in fact not be interpersonal communication but instead interaction with the shared digital artifact.

Structure and hierarchy In firms and other kinds of offline groups, one role of hierarchy is to coordinate actions and to make sure the organization is moving in the same direction (Coase, 1937). Even very large production-based online communities like Wikipedia almost completely lack a formal structure for assigning work: members work on what they want, when they want. Researchers have found that contributors often self-select into various "roles," performing actions like copy-editing, cleaning up after vandals, or welcoming newcomers (Welser et al., 2011). Although there are failures (Champion & Hill, 2021), this process works surprisingly well. One explanation is that the vast scale and low barriers to entry allow those with expertise to identify where improvements are needed and to make them (Benkler, 2002).

An important exception to the formal structurelessness of online communities is the role of moderators. One result of low barriers to entry and anonymity is that many online communities deal with many newcomers and bad-faith actors, and some community members act as moderators, who use powerful technical tools like bans and deletion of content. Often, moderator decisions are without recourse or appeal, and online communities can act as fiefdoms (Schneider, 2022). Even when moderators are chosen via more democratic means, those with the time and interest hold outsized power as measured by contributing to policy or having their contributions valued (Matei & Britt, 2017).

Of course, groups also undergo other emergent processes to shape norms and make decisions. Gibbs et al. (2021) argues that many online communities are able to exert "concertive control" on their members. The original theory of concertive control explained how in some conditions members of a work group will surveil each other, sanctioning norm violators and reinforcing rule-followers without management intervention (Barker, 1993). Gibbs et al. (2021) argue that the persistent visibility of interactions as well as technical tools like voting provide ways for online communities to develop and enforce norms even without strong interpersonal relationships or top-down moderation.

Algorithms and bots The communication and behavior of online community members is deeply influenced by the technical aspects of the platforms on which they reside. In particular, the algorithms driving and prioritizing certain posts and content over others can alter the context in which group communication happens and how it is perceived. For example, many communities include a voting system which automatically hides comments which have been downvoted by others or which comes from untrusted users (Lampe & Resnick, 2004). It is easy to see how systems like this shape which voices have influence.

Another distinct dynamic in online communities is the role of automated agents (i.e., bots). While there are some malignant bots, bots also act as beneficial, semi-visible group members, helping to moderate content, welcome newcomers, and enforce group norms (Seering et al., 2018). For example, a number of bots on Wikipedia identify and block vandals, fix typos, and alert contributors to possible problems (Geiger & Halfaker, 2013). By both contributing and shaping content in online communities, bots substantially change how communication flows within and across members of groups.

Scale The software and self-organizing processes of online communities allows them to exist across very different scales. While most communities are very small (Hwang & Foote, 2021), the same "space" can grow to accommodate hundreds of thousands of members (or more). How easily a group can scale depends on its goals and the software it uses. Voting-based conversation communities like Reddit may just need to add more moderators to handle a greater number of vandals while an open-source software project may need to add additional processes to manage contributions and ensure quality.

Output

Online communities frequently have outputs that are distinct from those of other types of groups. The tools of text-based, asynchronous communication which are the backbone of most online groups mean that many community outputs are also text-based, collaborative information goods, like wikis, open source software, or curated conversations. However, online communities also have relational and emotional outputs and produce impacts beyond their virtual spaces.

Information goods For some communities, the production of a shared artifact—a **public information good** (Fulk et al., 1996)—is the explicit goal. Fan communities on sites like Fandom, for example, collate information about the media content (e.g. television series or comics) they are fans of, recording detailed backstories and histories of characters as well as creating pages on other world-building aspects of the media such as fictional locations and creatures (Mittell, 2009). Other online communities facilitate the production of open source software, such as those on GitHub (Dabbish et al., 2012), or serve as important spaces for learning, such as those on StackOverflow or Reddit centered around specific skills like programming and design (Cheng et al., 2022).

Just like many face-to-face groups, some types of online communities do not produce clearly identifiable shared artifacts. However, online communities almost always produce a digital record of their interactions. In other words, as online communities archive informational exchanges, they inadvertently produce a persistent, searchable public informational archive that may benefit later viewers, in addition to any coordinated efforts that come out of the community.

Attachment and identity Distinct from the information or entertainment that online community artifacts and conversations provide, they can also produce social outputs just like many other kinds of groups, including support, camaraderie, and friendship (Ren et al., 2012). Because online community members are often strangers to one another and, in many cases, pseudoanonymous to each other, early news articles and books about online communities expressed shock and surprise that people could actually form meaningful relationships and a shared identity simply through text (Rheingold, 1993; Seabrook, 1998).

However, although dyadic friendships can be rare even in small online

communities (Hwang & Foote, 2021), people can still form a strong sense of group identity, trust each other, and find both informational and emotional support from groups of online strangers (Ren et al., 2012). As we hinted at earlier, there are aspects of online communities that may actually make forming a group identity easier. One is that groups are often focused on very specific and niche topics, which can appeal to individuals who already have a deep interest in the topic and are looking to find like-minded others. Other groups are explicitly centered on identity, such as the AAPI communities examined in Dosono and Semaan (2019). Second, the social identity model of deindividuation effects (SIDE) model suggests that pseudonymity and text-based communication can actually help people to focus less on the individual members of a group and to form a deeper relationship with the group as a whole (Reicher et al., 1995). Another possible explanation for the cohesion in some online communities is that the barriers to leaving are so low—those who disagree with a group decision or norm may just leave rather than creating schisms in the group (Hirschman, 1970).

Offline outcomes Online communities can influence people's beliefs and behaviors beyond interactions in the online space, including members' opinions on social issues as well as their willingness to participate in offline activism (Greijdanus et al., 2020). For example, Salehi et al. (2015) describes how Dynamo, a community platform designed to support Amazon Mechanical Turk workers, enabled community members to form publics around issues and mobilize collective campaigns to make their needs visible and to improve working conditions.

Unfortunately, online communities can also influence their members in destructive ways. Hannah (2021) argues that the QAnon conspiracy theory—which has led to multiple violent crimes—was enabled thanks in part to the way that anonymous online communities can provide legitimacy to conspiratorial thinking. More broadly, Massanari (2017) describes how the design, algorithms, and politics of a community platform can enable "toxic technocultures" that foster harassment campaigns including hate speech, doxxing, and threats of harm.

Implications for group communication

In summary, online communities are typically composed of an ever-changing group of pseudonymous strangers with little formal hierarchy or direction, embedded in a complex ecosystem of related communities. Researchers have

devoted significant attention to some aspects of how online community features influence group processes and group communication. For example, researchers have shown the importance of technological affordances in shaping how people in online groups communicate with one another (Kraut et al., 2012). At the same time, many aspects of group communication research have not been addressed explicitly and new theories and research are needed, especially around online community platforms as interdependent, self-organizing, multi-group systems. As we argue in the next section, the large-scale data from online communities is ideal for addressing this and other group communication topics.

Data from online communities

In addition to being novel empirical settings for studying groups, online communities are also exciting to researchers due to the type and scale of the data they produce. Simply as part of their operation, online platforms like Wikipedia, Reddit, and Github store billions of timestamped actions and interactions. As discussed earlier, these actions can be made visible and can contribute to the informational public goods produced by a community. They can also be used to study social science questions: As actions like joining a new community or sending a message to another user are naturally archived, they become **digital trace data** that can be both qualitatively and quantitatively analyzed to study social behaviors in these spaces.

What is so special about online data? Among other attributes, online data is often large-scale, longitudinal, and granular. Large-scale can be massive. For example, publicly available Reddit comment data includes *billions* of comments in *millions* of communities. This is especially exciting for group researchers, because it is possible to study large numbers of groups at once, to study groups that might otherwise be too small to meaningfully identify and sample from (Welles, 2014), and to treat groups as the unit of analysis. Online community data is also longitudinal, meaning that it's tracked over time. In most systems, data gathering is "always on"—actions are tracked constantly over extended periods. Finally, online data is often very granular, with full-text data that can be tied to individual users, letting researchers study within-group and within-individual processes.

Of course, there are also some characteristics of online community data which are problematic either for researchers or for community members. In the following section, we discuss some of the ethical considerations, and later

on we discuss some of the technical difficulties.

Ethical questions in researching online communities

Although large-scale online community data presents exciting opportunities for social science research, it also poses ethical questions around privacy and consent.

While the fine-grained nature of some digital trace data is a boon for researchers, it raises privacy concerns at both the community level and the individual level. Technically, the content of many online communities is public, and it could be argued that research using this data is similar to researchers observing people in a public park. However, norms around how to do research using this data (and whether researchers even *should* use this data) are evolving (Fiesler & Proferes, 2018; Hallinan et al., 2020).

One example of these debates focuses on a method of data collection called web scraping. Through web scraping, a researcher uses a custom code script to access and collect anything that they could access through a browser, including private forums or personal information from friends. The scale and reach of this automated process makes it much harder to argue that it is analogous to observing a public space. These concerns become amplified when researchers are studying vulnerable or marginalized communities where unwanted attention can put members at risk. On the other hand, web scraping can also be a powerful tool to audit how online community platforms' algorithms and design choices affect their users. Through web scraping, researchers can study aspects of corporate platforms that corporations would rather keep hidden (Bandy, 2021).

When it comes to consent, online data gathering is equally fraught. Technically, when people create accounts on platforms, they consent to having their content made public. However, online communities are not spaces intended for research: many community members participate in communities with no idea that their clicks, likes, and comments might be used as data. Some community members may feel that their activity being used as data affects the integrity of their communal space, especially if the topic being discussed is sensitive; others may feel like the community is being exploited for research purposes, especially when the research involves experiments that manipulate user experiences. An infamous case is the 2014 study on emotional contagion on Facebook, where researchers manipulated which posts appeared in users' newsfeeds to see whether users would post more positively

or negatively after being exposed to more or less positive content (Kramer et al., 2014). In addition to the ethical implications of attempting to manipulate peoples' emotions, the experiment caused an uproar because none of the users were not aware that they had been experimented on until the study was published.

In short, research using online community data has enormous promise, but also new challenges and opportunities for harm that do not always have clear-cut answers on how to proceed. This chapter covers just the tip of the iceberg as an introduction to studying online communities, but it is imperative that any researcher of online communities carefully considers what the relationship of their research is with the communities being studied. Researchers should seek to increase the benefits of research to communities—for example, by conducting participatory research *alongside* community members, in order to explicitly help communities meet their goals (Matias, 2019). They should also seek to reduce the potential for harms by obfuscating details of participants and their communities (even for public accounts and public data), reporting data in aggregate, and working with community members and other researchers to think through the implications of proposed research (Vitak et al., 2016). When done well, online community research can yield findings that are ultimately helpful for society and for online communities.

Online community research

In this section, we give examples of how large-scale data can be used to research online communities. Much of the research on online communities does not require large-scale data; methods like surveys, participant observation, experiments, interviews, and ethnography have generated much of our understanding of how and why online communities work. However, because large-scale online datasets are opening up new avenues of inquiry, we focus on a few exciting approaches and opportunities enabled by them.

Observational studies The first approach is observational studies. This typically involves gathering digital trace data about individuals or groups, creating measures from that data that correspond to theoretical constructs, and then using statistical analyses to test hypotheses about the relationship between those constructs. Researchers can focus on individuals, groups, or ecosystems of groups. Work at the individual level looks at how members interact with and are influenced by a group that they belong to. For example, Danescu-Niculescu-Mizil et al. (2013) use natural language-processing (NLP)

tools to look at the longitudinal dynamics of beer-rating online communities. They show that the community's linguistic norms changed over time, and that newer members of a community were more willing to adapt to linguistic shifts (e.g., using 'aroma' instead of 'smell'), while older members left when the norms changed too much.

At the group level, researchers have looked at things like how the structure of the communication networks in a group predict the group's longevity or productivity (Foote et al., 2023). Comparing across groups can be powerful. For example, TeBlunthuis et al. (2018) look at how hundreds of popular wikis changed in size over time; they find a general pattern where wikis would grow for a few years followed by a gradual decline in activity. Researchers have also begun examining "ecosystems" of communities. Given how easy it is for people to move between communities or participate in multiple communities, it can be illuminating to study the relationships that communities have with each other. For example, researchers showed that the amount of overlap that a community has with other communities—either in membership or topic—has an inverse-U (⌒) relationship with the community's activity level and survival (Zhu et al., 2014).

Natural experiments One special case of observational research is natural experiments. In a natural experiment, researchers look for times when external, unexpected "shocks" impact a system. Then, due to the "always on" nature of online data, researchers go back in time to look at the influence of the shock. For example, Zhang and Zhu (2011) identified a time when the Chinese government blocked Wikipedia for nearly a year without warning, dramatically reducing the size of the Chinese Wikipedia community. According to some theories, as group size grows so does the temptation to free ride, so we would expect people to increase their contributions when the community shrank (Olson, 1965). On the other hand, perhaps people contribute partially because it feels good to help others (the "warm glow theory"), and so a larger group of readers and co-contributors would encourage one to contribute more (Andreoni, 1990). Zhang and Zhu (2011) found that individuals who were active before the block (but not blocked themselves) actually contributed less during the block, providing evidence for the "warm glow theory" that people are motivated in part by the knowledge that they are helping others.

Experiments Online data can also be leveraged for participants doing "real" experiments. For example, Matias (2019) worked with moderators of the very

popular 'r/science' subreddit to show a "stickied" comment containing the community rules at the top of a random selection of posts. Newcomers who participated on those posts that had a comment were both more likely to participate and were more likely to communicate according to group norms. This kind of approach enables researchers to create large-scale experiments at a much lower cost than lab-based experiments. In the case of the Reddit experiment, Matias (2019) studied nearly 63,000 newcomer participants. Always-on data also means that the behavior of those in experiments can be tracked both before and after the experiment, allowing for longitudinal and long-term analyses.

Computational text analysis Online data often includes the text of thousands or millions of interactions. Qualitative methods like content analysis or ethnography are designed to make meaning out of texts, but when dealing with the equivalent of hundreds of thousands of pages of text, these methods are impossible. Instead, researchers often use natural language processing (NLP) tools which use computation to summarize text in some way. LIWC is a popular utility that defines a list of terms that correspond to psychological constructs and counts how often they occur in a text or set of texts (Tausczik & Pennebaker, 2010). For example, Hamilton et al. (2017) showed that Reddit users who used more personal pronouns and affect words were less likely to quit a community. More advanced NLP approaches include topic modeling, which seeks to recover different "topics" that are used in large sets of texts, based on how often words co-occur within texts (Blei, 2012). Some recent research seeks to build best-of-both-worlds processes that combine automated steps done by a computer and interpretive steps done by humans. Nelson (2020) suggests "Computational Grounded Theory," an approach which uses topic modeling or other natural language processing steps to identify patterns in the data and to identify texts which are representative of those patterns. The next step involves a "computationally-guided deep reading," intended to contextualize, question, and interpret the key texts identified in the first step. These two inductive steps help to build theories which can then be tested using other computational methods (like LIWC). Ideally, this approach builds on the strengths of both humans and computers and allows for a rich understanding of even very large datasets.

EVIDENCE-BASED RECOMMENDATIONS FOR RESEARCHERS

Each of the research approaches that we have outlined opens up exciting new opportunities for studying group communication phenomena by taking advantage of the scale and granularity of online community data. There are, of course some drawbacks to working with online data. In addition to some of the ethical and conceptual difficulties discussed above, working with this type of data also requires computational and statistical training.

For some datasets, even obtaining and storing the data requires significant computational expertise. In order to gather a sufficient longitudinal dataset, researchers must consistently and securely collect, clean, and store data in a reliable manner. In some cases, platforms make large datasets available for download, but often researcher must write and maintain code that runs for an extended period of time on a server. Researchers benefit from developing well-documented data collection pipelines. Doing so can also strengthen the robustness and validity of one's work and analyses because it can enable replication as well as visibility into the strengths and limitations of the data collection process.

More broadly, there are a number of challenges when working with large-scale online data, including ensuring a match between data measures and theoretical constructs, successfully applying approaches like large-scale natural experiments or NLP, and accounting for changes to online platforms (Salganik, 2017); each of these can require fairly advanced computational skills. For example, one common challenge is detecting bots in digital trace data. If a researcher wants to know about human behaviors, it's important to distinguish which posts and comments are by humans and which are by bots. While some bots are obviously labeled or detectable because of behavioral patterns (posting too fast, always writing the same message, etc.), some are not. Failing to sufficiently account for non-human contributions can paint a misleading picture of how people interact online. Because of the scale of the data, filtering out bots can require researchers to take approaches like building or applying classification algorithms to predict if a user in their data is a bot.

Fortunately, many of the skills needed can be learned in a few semesters through publicly available learning resources or through tailored courses. Alternatively, this area of work offers an exciting opportunity for researchers to build interdisciplinary collaborations with computer scientists.

Conclusion

Online communities have emerged as an important new way of organizing groups. From business to politics to culture, online communities are increasingly influencing how people perceive and act in the world: they produce powerful and popular software, gather and produce knowledge, and organize and persuade. Because groups are digital—with interactions that are timestamped and stored—they can be studied with more depth than we can study groups in other contexts. Online communities offer an incredible chance to understand how communication technologies are transforming group communication in evolving empirical settings as well as to re-visit fundamental questions about how individuals in groups can effectively interact, organize, and communicate with one another. Only through understanding the dynamics of online communities better will we be able to shape the role that they play in society.

Box: The effect of community bans

While we have focused this chapter on the positive aspects of online communities, many individuals and entire communities engage in racism, political extremism, misogyny, or bigotry. One common approach platforms take to dealing with problematic communities is a ban, when the community space is removed from the platform.

Does this work, or do members of a banned community simply move to different communities and behave in the same way? In 2015, Reddit change their platform policies and unexpectedly banned a large number of their most troublesome subreddits. Chandrasekharan et al. (2017) looked at what happened, finding that the ban was effective: users who had been active in toxic communities and stayed on Reddit reduced their use of toxic language dramatically, and there was not a significant uptick in toxic language in the other communities that they joined.

This is great news for platforms, but we might ask the same question at the platform level: if a community is banned, do its users simply migrate to a new platform? Ribeiro et al. (2021) studied two cases where users from banned communities created their own new standalone sites. They found that these new communities were much less popular, with fewer users, posts, and new recruits. However, in one of the two cases studied users on the new platform showed increased linguistic markers of toxicity and radicalization.

There are a few lessons that we might take from this research, some of which echo arguments from this chapter. One lesson is the power of sociotechnical tools: banning turned out to be a simple but effective tool. Another lesson is the importance of group norms to shape behavior; users who behaved badly in a toxic community changed their behavior when participating in spaces with more prosocial norms.

Finally, there seems to be a lesson about the importance of platforms of related communities. If we imagine online communities like digital "clubs" with meet in digital "rooms," a ban is like kicking a club out of its meeting place. Even though finding a new digital place to meet seems relatively cost-free, when a community is banned, the "club" typically disbands and only a fraction of its users find a way to coalesce in a new space.

Additional Readings

Bruckman, A. S. (2022). *Should you believe wikipedia?: Online communities and the construction of knowledge*. Cambridge University Press
Kraut, R. E., Resnick, P., & Kiesler, S. (2012). *Building successful online communities: Evidence-based social design*. MIT Press
Salganik, M. J. (2017). *Bit by Bit: Social Research in the Digital Age*. Princeton University Press

Key Words

- Digital trace data - As people participate in online communities, data and metadata about what they are doing is stored. This data is often visible to other users and/or made available to researchers.
- Lurkers - People who consume the content on a community but don't actively participate. For most online communities, lurkers represent the vast majority of community members.
- Pseudonymous - Many online communities allow for pseudonyms—persistent identifiers like usernames which are not tied to a users real identity.
- Public information good - A shared information repository. Many online communities produce explicit information goods, like wikis or software. Stored conversations can also serve as public information goods.
- Stigmergic communication - communication that happens through modifying the environment rather than through typical communication channels.

Abstract Online technologies allow people to create and participate in online groups called online communities. These groups have a number of differences and similarities with traditional face-to-face groups and virtual work teams, including differences in who participates, the communicative and technological tools used, and the goals of these communities. Despite being composed of pseudonymous volunteers, online communities can coordinate work, create group identity, and develop shared norms. Because online community platforms track the behavior of group members and store these in large-scale, longitudinal databases, researchers can study them using new approaches and can ask new questions. This provides a unique opportunity for understanding dynamics of groups such as how online communities and groups form, how people and group members they change over time, and how groups relate to and interact with one another.

References

Ammari, T., Schoenebeck, S., & Romero, D. (2019). Self-declared throwaway accounts on reddit: How platform affordances and shared norms enable parenting disclosure and support. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 135:1–135:30. https://doi.org/10.1145/3359237

Andreoni, J. (1990). Impure Altruism and Donations to Public Goods: A Theory of Warm-Glow Giving. *The Economic Journal*, *100*(401), 464–477.

Bandy, J. (2021). Problematic Machine Behavior: A Systematic Literature Review of Algorithm Audits. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW1), 74:1–74:34. https://doi.org/10.1145/3449148

Barker, J. R. (1993). Tightening the Iron Cage: Concertive Control in Self-Managing Teams. *Administrative Science Quarterly*, *38*(3), 408–437. https://doi.org/10.2307/2393374

Benkler, Y. (2002). Coase's penguin, or, Linux and the nature of the firm. *Yale Law Journal*, *112*(3), 369–446. https://doi.org/10.2307/1562247

Blei, D. M. (2012). Probabilistic Topic Models. *Commun. ACM*, *55*(4), 77–84. https://doi.org/10.1145/2133806.2133826

Bruckman, A. S. (2022). *Should you believe wikipedia?: Online communities and the construction of knowledge*. Cambridge University Press.

Champion, K., & Hill, B. M. (2021). Underproduction: An approach for measuring risk in open source software. *2021 IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER)*, 388–399. https://doi.org/10.1109/SANER50967.2021.00043

Chandrasekharan, E., Pavalanathan, U., Srinivasan, A., Glynn, A., Eisenstein, J., & Gilbert, E. (2017). You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech. *Proc. ACM Hum.-Comput. Interact.*, *1*(CSCW), 31:1–31:22. https://doi.org/10.1145/3134666

Cheng, R., Dasgupta, S., & Hill, B. M. (2022). How Interest-Driven Content Creation Shapes Opportunities for Informal Learning in Scratch: A Case Study on Novices' Use of Data Structures. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 1–16. https://doi.org/10.1145/3491102.3502124

Coase, R. H. (1937). The Nature of the Firm. *Economica*, *4*(16), 386–405. https://doi.org/10.1111/j.1468-0335.1937.tb00002.x

Dabbish, L., Stuart, C., Tsay, J., & Herbsleb, J. (2012). Social coding in GitHub: Transparency and collaboration in an open software repository. *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work - CSCW '12*, 1277. https://doi.org/10.1145/2145204.2145396

Danescu-Niculescu-Mizil, C., West, R., Jurafsky, D., Leskovec, J., & Potts, C. (2013). No country for old members: User lifecycle and linguistic change in online communities. *Proceedings of the 22nd International Conference on World Wide Web - WWW '13*, 307–318. https://doi.org/10.1145/2488388.2488416

Dosono, B., & Semaan, B. (2019). Moderation practices as emotional labor in sustaining online communities: The case of aapi identity work on reddit. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3290605.3300372

Fiesler, C., & Proferes, N. (2018). "Participant" Perceptions of Twitter Research Ethics. *Social Media + Society*, *4*(1), 2056305118763366. https://doi.org/10.1177/2056305118763366

Foote, J. (2022). A systems approach to studying online communities. *Media and Communication*, *10*(2), 29–40. https://doi.org/10.17645/mac.v10i2.5042

Foote, J., Shaw, A., & Hill, B. M. (2023). Communication networks do not predict success in attempts at peer production. *Journal of Computer-Mediated Communication*, *28*(3), zmad002. https://doi.org/10.1093/jcmc/zmad002

Fulk, J., Flanagin, A. J., Kalman, M. E., Monge, P. R., & Ryan, T. (1996). Connective and communal public goods in interactive communication systems. *Communication Theory*, *6*(1), 60–87. https://doi.org/10.1111/j.1468-2885.1996.tb00120.x

Geiger, R. S., & Halfaker, A. (2013). When the levee breaks: Without bots, what happens to Wikipedia's quality control processes? *Proceedings of the 9th International Symposium on Open Collaboration (OpenSym '13)*, 6:1–6:6. https://doi.org/10.1145/2491055.2491061

Gibbs, J. L., Rice, R. E., & Kirkwood, G. L. (2021). Digital discipline: Theorizing concertive control in online communities. *Communication Theory*. https://doi.org/10.1093/ct/qtab017

Greijdanus, H., de Matos Fernandes, C. A., Turner-Zwinkels, F., Honari, A., Roos, C. A., Rosenbusch, H., & Postmes, T. (2020). The psychology of online activism and social movements: Relations between online and offline collective action. *Current Opinion in Psychology*, *35*, 49–54. https://doi.org/10.1016/j.copsyc.2020.03.003

Hallinan, B., Brubaker, J. R., & Fiesler, C. (2020). Unexpected expectations: Public reaction to the Facebook emotional contagion study. *New Media & Society*, *22*(6), 1076–1094. https://doi.org/10.1177/1461444819876944

Hamilton, W. L., Zhang, J., Danescu-Niculescu-Mizil, C., Jurafsky, D., & Leskovec, J. (2017). Loyalty in online communities. *arXiv:1703.03386 [cs]*.

Hannah, M. (2021). QAnon and the information dark age. *First Monday*. https://doi.org/10.5210/fm.v26i2.10868

Heylighen, F. (2016). Stigmergy as a universal coordination mechanism I: Definition and components. *Cognitive Systems Research*, *38*, 4–13. https://doi.org/10.1016/j.cogsys.2015.12.002

Hirschman, A. O. (1970). *Exit, Voice, and Loyalty: Responses to Decline in Firms, Organizations, and States*. Harvard University Press.

Hwang, S., & Foote, J. (2021). Why do people participate in small online communities? *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW2), 462:1–462:25. https://doi.org/10.1145/3479606

Ilgen, D. R., Hollenbeck, J. R., Johnson, M., & Jundt, D. (2005). Teams in Organizations: From Input-Process-Output Models to IMOI Models. *Annual Review of Psychology*, *56*(1), 517–543. https://doi.org/10.1146/annurev.psych.56.091103.070250

Kramer, A. D. I., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, *111*(24), 8788–8790. https://doi.org/10.1073/pnas.1320040111

Kraut, R. E., Resnick, P., & Kiesler, S. (2012). *Building successful online communities: Evidence-based social design*. MIT Press.

Lampe, C., & Resnick, P. (2004). Slash(dot) and burn: Distributed moderation in a large online conversation space. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 543–550. https://doi.org/10.1145/985692.985761

Lampe, C., Wash, R., Velasquez, A., & Ozkaya, E. (2010). Motivations to participate in online communities. *Proceedings of the 28th International Conference on Human Factors in Computing Systems*, 1927–1936. https://doi.org/10.1145/1753326.1753616

Massanari, A. (2017). #Gamergate and The Fappening: How Reddit's algorithm, governance, and culture support toxic technocultures. *New Media & Society*, *19*(3), 329–346. https://doi.org/10.1177/1461444815608807

Matei, S. A., & Britt, B. C. (2017). *Structural differentiation in social media: Adhocracy, entropy, and the "1 % effect"*. Springer.

Matias, J. N. (2019). Preventing harassment and increasing group participation through social norms in 2,190 online science discussions. *Proceedings of the National Academy of Sciences*, *116*(20), 9785–9789. https://doi.org/10.1073/pnas.1813486116

Mittell, J. (2009). Sites of participation: Wiki fandom and the case of Lostpedia. *Transformative Works and Cultures*, *3*. https://doi.org/10.3983/twc.2009.0118

Nelson, L. K. (2020). Computational Grounded Theory: A Methodological Framework. *Sociological Methods & Research*, *49*(1), 3–42. https://doi.org/10.1177/0049124117729703

Nonnecke, B., & Preece, J. (2000). Lurker demographics: Counting the silent. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 73–80. https://doi.org/10.1145/332040.332409

Olson, M. (1965). *The logic of collective action: Public goods and the theory of groups*. Harvard University Press.

Phillips, M., Lorenz, T., Bernard, T. S., & Friedman, G. (2021). The Hopes That Rose and Fell With GameStop. *The New York Times*.

Reicher, S. D., Spears, R., & Postmes, T. (1995). A Social Identity Model of Deindividuation Phenomena. *European Review of Social Psychology*, *6*(1), 161–198. https://doi.org/10.1080/14792779443000049

Ren, Y., Harper, F., Drenner, S., Terveen, L., Kiesler, S., Riedl, J., & Kraut, R. (2012). Building member attachment in online communities: Applying theories of group identity and interpersonal bonds. *Management Information Systems Quarterly*, *36*(3), 841–864. https://doi.org/10.2307/41703483

Rheingold, H. (1993). *The virtual community: Homesteading on the electronic frontier*. Addison Wesley Publishing Company.

Ribeiro, M. H., Jhaver, S., Zannettou, S., Blackburn, J., Stringhini, G., De Cristofaro, E., & West, R. (2021). Do Platform Migrations Compromise Content Moderation? Evidence from r/The_Donald and r/Incels. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW2), 316:1–316:24. https://doi.org/10.1145/3476057

Ruggiero, T. E. (2000). Uses and gratifications theory in the 21st century. *Mass Communication and Society*, *3*(1), 3–37. https://doi.org/10.1207/S15327825MCS0301_02

Salehi, N., Irani, L. C., Bernstein, M. S., Alkhatib, A., Ogbe, E., Milland, K., & Clickhappier. (2015). We are dynamo: Overcoming stalling and friction in collective action for crowd workers. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 1621–1630. https://doi.org/10.1145/2702123.2702508

Salganik, M. J. (2017). *Bit by Bit: Social Research in the Digital Age*. Princeton University Press.

Schneider, N. (2022). Admins, mods, and benevolent dictators for life: The implicit feudalism of online communities. *New Media & Society*, *24*(9), 1965–1985. https://doi.org/10.1177/1461444820986553

Seabrook, J. (1998). *Deeper*. Simon and Schuster.

Seering, J., Flores, J. P., Savage, S., & Hammer, J. (2018). The social roles of bots: Evaluating impact of bots on discussions in online communities.

Proc. ACM Hum.-Comput. Interact., *2*(CSCW), 157:1–157:29. https://doi.org/10.1145/3274426

Suler, J. (2004). The online disinhibition effect. *CyberPsychology & Behavior*, *7*(3), 321–326. https://doi.org/10.1089/1094931041291295

Tausczik, Y. R., & Pennebaker, J. W. (2010). The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, *29*(1), 24–54. https://doi.org/10.1177/0261927X09351676

TeBlunthuis, N., Kiene, C., Brown, I., Levi, L. (, McGinnis, N., & Hill, B. M. (2022). No Community Can Do Everything: Why People Participate in Similar Online Communities. *Proceedings of the ACM on Human-Computer Interaction: Computer Supported Cooperative Work*, *6*, 1–25. https://doi.org/10.1145/3512908

TeBlunthuis, N., Shaw, A., & Hill, B. M. (2018). Revisiting "The rise and decline" in a population of peer production projects. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 355:1–355:7. https://doi.org/10.1145/3173574.3173929

Treem, J. W., & Leonardi, P. M. (2013). Social media use in organizations: Exploring the affordances of visibility, editability, persistence, and association. *Annals of the International Communication Association*, *36*(1), 143–189. https://doi.org/10.1080/23808985.2013.11679130

Vitak, J., Shilton, K., & Ashktorab, Z. (2016). Beyond the Belmont Principles: Ethical Challenges, Practices, and Beliefs in the Online Data Research Community. *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, 941–953. https://doi.org/10.1145/2818048.2820078

Wegner, D. M. (1987). Transactive Memory: A Contemporary Analysis of the Group Mind. In B. Mullen & G. R. Goethals (Eds.), *Theories of Group Behavior* (pp. 185–208). Springer. https://doi.org/10.1007/978-1-4612-4634-3_9

Welles, B. F. (2014). On minorities and outliers: The case for making Big Data small. *Big Data & Society*, *1*(1), 2053951714540613. https://doi.org/10.1177/2053951714540613

Welser, H. T., Cosley, D., Kossinets, G., Lin, A., Dokshin, F., Gay, G., & Smith, M. (2011). Finding social roles in Wikipedia. *Proceedings of the 2011 iConference*, 122–129. https://doi.org/10.1145/1940761.1940778

Zhang, X. M., & Zhu, F. (2011). Group size and incentives to contribute: A natural experiment at chinese wikipedia. *American Economic Review*, *101*(4), 1601–1615. https://doi.org/10.1257/aer.101.4.1601

Zhu, H., Chen, J., Matthews, T., Pal, A., Badenes, H., & Kraut, R. E. (2014). Selecting an effective niche: An ecological view of the success of online

communities. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 301–310. https://doi.org/10.1145/2556288.2557348