CESifo WORKING PAPERS

10841
2023

December 2023

The Heterogeneous Effect of Digitizing Community Activities on Community Participation

Martina Pocchiari, Jason M.T. Roos

CESifo

Impressum:

CESifo Working Papers
ISSN 2364-1428 (electronic version)
Publisher and distributor: Munich Society for the Promotion of Economic Research - CESifo GmbH
The international platform of Ludwigs-Maximilians University's Center for Economic Studies and the ifo Institute
Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740, Telefax +49 (0)89 2180-17845, email office@cesifo.de
Editor: Clemens Fuest
https://www.cesifo.org/en/wp
An electronic version of the paper may be downloaded
· from the SSRN website: www.SSRN.com
· from the RePEc website: www.RePEc.org
· from the CESifo website: https://www.cesifo.org/en/wp

CESifo *Working Paper No. 10841*

The Heterogeneous Effect of Digitizing Community Activities on Community Participation

Abstract

As social and business activities undergo significant digital transformation, the allure of digitized versus traditional in-person events remains a critical question. This study quantifies the heterogeneous effect of digitization on community participation across communities catering to different interests, such as business, technology, and socialization. Leveraging comprehensive panel data from a leading community-building platform, our analyses reveal that prospective participation in digitized events is lower compared to their in-person counterparts. However, this overarching effect masks a more complex reality. We uncover significant heterogeneity in participation intentions, with outcomes closely tied to specific event characteristics and interest topics. For example, prospective engagement in digitized networking meetings varies substantially depending on whether the meetings are business- or leisure-oriented. Similarly, the context in which goal-oriented meetings occur plays a decisive role: prospective participation in digitized events decreases by 2.95% for writing clubs, but increases by 2.72% for language clubs. These findings suggest that approaches to community-building through digitization must be tailored to the unique attributes of the community and its events. A one-size-fits-all strategy to digital transformation may fall short of fostering meaningful engagement. As the digital landscape evolves, our research offers quantitative benchmarks and managerial guidelines for community engagement in the digital age.

JEL-Codes: M310, M100.

*Martina Pocchiari**
NUS Business School
National University of Singapore
pmartina@nus.edu.sg

Jason M.T. Roos
Rotterdam School of Management
Erasmus University Rotterdam / NL
jroos@rsm.nl

*corresponding author

This version: August 2023
The authors acknowledge the support of all those who helped improving this study over time. We thank the research assistants, the faculty and PhD candidates of the Marketing Management department at the Rotterdam School of Management, and Francesco Capozza for their support and feedback. We also thank the faculty and PhD students of the Doctoral Colloquium of EMAC 2019-2021, the Doctoral Consortium of Marketing Science 2020, and the Doctoral Colloquium of the Italian Marketing Society 2020 for their comments. We thank Max Pachali, Andreas Bayerl, and JK Chong for their helpful feedback. Finally, we thank the community managers who shared their expert insights on community digitization with us: Anastasia "Stacy" Raspopina, Tristan Lombard, Scott Tran, Tali Vasilevsky, Samy Zerrouki, and the hosts and audience of the Communities Show.

1 Introduction

Digitization technologies enable social and business activities to take novel formats, such as livestreams, webcasts, online meetings, and asychnronous media feeds. These novel formats often appear more affordable, accessible, and efficient than in-person alternatives. As a result, many events – such as brand-building conferences, product workshops and demonstrations, brand community events, social support groups, and sports classes – have migrated from face-to-face to digital formats in recent years. The COVID–19 pandemic has expedited this ongoing trend towards digitization.

In addition to lowering organization and participation costs, digitized activities present new opportunities for brand and community development, and for engaging people at a large scale. But the potential to foster engagement is not uniform. On the one hand, for example, Lee et al. (2022) and Sconti (2022) suggest that digitizing courses may improve academic involvement in the context of higher education. Similarly, Riedl et al. (2021) find that virtual meetings for co-working teams can enable relationship-building, if team members are socially receptive. On the other hand, Cohn et al. (2022) and Touré-Tillery and Wang (2022) suggest that digitized social interactions make people less prone to prosociality; and Aarons-Mele (2022) and Capossela (2022) report that online meetings and remote work may discourage employee engagement, and make employees more anxious and more desirous of intentional social connections. Business leaders are therefore challenged with balancing digitized, hybrid, and in-person formats for their events, and in order to improve participation must determine which activities offer more value if digitized (Bonsall 2022, Gratton 2021, Yokoi et al. 2021).

Marketers and business leaders are not only confronted with uncertainty about how well digitized activities can engage different audiences, they also lack insights into the effects of digitization across different types of activities and contexts. For example, a product manager deciding whether to organize technical training in-person or on-line would find conflicting results in the literature, as customers' learning experience may be either enhanced or hindered by virtual formats (Bettinger et al. 2017, Lee et al. 2022, Sconti 2022). Moreover, extant studies typically inform only certain types of organizations and brands. For instance, the aforementioned studies were conducted on undergraduate students at a large for-profit university (Bettinger et al. 2017), on Korean students of social studies and science classes at selected schools (Lee et al. 2022), and on Italian high school students participating in a financial education program (Sconti 2022). At present, there are no large-scale studies of event digitization involving many organizations, communities, or companies, and spanning a wide range of interest categories, to inform managers about the likely effects of digitization.

To fill this gap, the present study quantifies the heterogeneous impact of digitizing social activities on prospective attendee participation across activities catering to different interests, including business, education, entertainment, and socialization. We do this using a unique and highly detailed panel data set collected from Meetup, a prominent global platform for organizing shared-interest communities and social activities. The data describe 118,326 events organized by 12,132 communities (called Meetup *groups*) in the first half of 2019 – before the Covid–19 pandemic forced many community events to become digitized. Focusing on the pre-Covid period allows us to study the impact of digitization in the absence of exceptional discomfort and concern about in-person interactions.

The estimation data include details about the groups, their events, their members, and their members' participation in those events. As the data do not include a "digitization" variable, we train Support Vector Machines (SVMs) on a subset of free-text event descriptions to impute the degree of digitization of each event. Event digitization is thus measured as a continuous probability that an event is digitized, as predicted by the SVMs. To our knowledge, this is the first study to measure the degree of digitization of social events using unstructured text data. RSVPs of community members provide a measure of their intentions to participate in digital and in-person events, and serve as the outcome of interest in this study.

Identifying the causal impact of event digitization on members' participation is not trivial in this setting. The main challenge to identification arises from the potential for unobservable factors to influence both the extent of event digitization and members' participation decisions. To address this, we employ a set of relevant covariates and fixed effects that vary by group, event, member, and time. We thereby control for the market-level and group-level demand for digitized events. Controlling for these factors allows us to leverage the panel structure of the data to estimate causal effects, with the effect of digitization identified by repeated, within-member exposure to both digitized and in-person events.

To estimate the effect of activity digitization on participation, we specify a Bayesian structural causal model (SCM) that allows for heterogeneity in the effect of digitization across groups catering to different interest categories. The parameter estimates yield two important insights. First, overall, people participate less in digitized events compared to similar in-person events. However, the effect of digitizing events varies significantly across groups catering to different interest topics. We find that digitization lowers participation in events related to socializing and matchmaking, social support and social causes (support among women, local politics, environment), sports and photography, language courses and public speaking, and spirituality. In contrast, digitization increases participation in events related to business (business network-

ing, innovation and start-ups, career growth), technology (software development, coding, web design), financial investments and real estate, and gaming. Anecdotal evidence provided to business leaders suggests that digitization may be more suitable for events with specific characteristics, such as the extent of interactivity (Clark 2021) and the nature of the event goals (e.g., relationship- versus task-based, Ringel 2021). However, our results indicate that even events with similar features may not be equally suitable for digitization, if they are catering to different interest topics. For example, both socialization and business networking events require high levels of social interaction. Similarly, both a language class and a coding workshop mostly pursue task-based goals. However, the effect of digitization on prospective attendance to one type of event is almost opposite to the other.

To provide more precise managerial recommendations, we perform a counterfactual policy evaluation. The simulated policy entails switching digitized events in the sample from fully in-person to fully digitized formats, and measuring changes to RSVP decisions due to digitization. The analysis, therefore, yields conditional average treatment effects on the treated (CATTs) across groups and interest topics. Focusing the counterfactual analysis on treated groups has two main advantages. First, most digitized events could be shifted to in-person formats without fundamentally altering their characteristics, while the opposite is not necessarily possible. For example, it is easier to imagine shifting a coding workshop to an in-person format without deviating from its goals and properties, than shifting a beer tasting to an online setting. Second, while attendees of in-person events may have never attended a virtual meeting, attendees to digitized events are more likely to have some experience with in-person events. Focusing on treated groups, therefore, lends pragmatism to a discussion of counterfactual RSVP decisions.

The counterfactual analysis allows us to disentangle the effect of digitization on different forms of participation, and in particular on the extent to which members express positive intentions to participate. For instance, we show that digitization causes a decrease in positive response rates to "Writing and Hobbies" events, and an increase in positive responses to "Business Networking" events. The counterfactual analysis also provides additional quantitative insight into the heterogeneous, nuanced effect of digitization on community engagement. For example, in counterfactual scenarios, digitization can have opposite effects when goal-oriented meetings are organized in different contexts: average participation in online writing clubs decreases by 2.95%, while average participation in language clubs increases by 2.72%.

We test the robustness of the structural causal results in two ways. First, we estimate the average treatment effect of digitization among treated individuals using non-parametric causal random forests (CRFs; Athey and Wager 2019, Wager and Athey 2018). We use random forests

to relax the parametric functional form assumptions of the structural causal model. This robustness check reveals that the heterogeneous treatment effects estimated in the counterfactual analysis are overall robust to alternative model specifications. Second, we perform nearest-neighbor matching on observable covariates, and estimate group-level CATTs. Analyses based on the matched data address potential concerns about covariate imbalance between digitized and in-person events, and show that covariate balance does not fundamentally alter the counterfactual results. Furthermore, insights from the nearest-neighbor matching model suggest that the main results are robust to a binary specification of the digitization variable.

These results contribute to several areas in the literature on digitization. First, this study contributes to the literature on the effects of digitization on economic behavior, such as cooperation, coordination, and contribution to a public good (e.g., Bettinger et al. 2017, Bourreau and Doğan 2018, Cohn et al. 2022). Past studies have investigated the role of digitizing human experiences in controlled lab experiments, or in narrow empirical settings (e.g., coin-tossing tasks, CD gift-giving, education), and have provided important initial evidence on the impact of digitized interactions on social engagement. We contribute to this work by considering the impact of digitization in the field, using a sample of thousands of social groups and events across a wide range of interest categories.

Second, our work is related to studies of the impact of digitization on community success. Previous studies have considered the effect of increased community digitization in the context of individual communities that were organized either completely online or completely in-person (Algesheimer et al. 2010, Dessart et al. 2015, Kang et al. 2014, Wiertz and de Ruyter 2007). The present study complements this literature by assessing how digitized interactions affect community participation across events with varying degrees of digitization, while controlling for the characteristics of communities, members, and events. Finally, the study also contributes to the stream of literature studying the differential impact of physical versus digital experience formats on customer intentions (e.g., Atasoy and Morewedge 2018, Catapano et al. 2022, Luangrath et al. 2022, Touré-Tillery and Wang 2022, Wiegand and Imschloss 2021). These studies have investigated customers' intentions to contribute to charities, preferences for entertainment services in hypothetical scenarios, and intentions to purchase automotive products. We complement this line of work with insights about additional types of products and experiences subject to various degrees of digitization, including fitness classes, panel discussions, information sessions, social support groups, and language lessons.

The results of this study also have important implications for decision makers responsible for local or distributed communities, marketing events, neighborhood groups, workforces, and

educational groups. The Covid–19 pandemic led to a surge in the demand for digitizing shared-interest communities. As such, investing in the digitization of customer experiences has become more crucial than ever before (The CMO Survey 2021). This study provides novel insights for these stakeholders. First, event marketers who strive to maximize event engagement can better understand which types of events are most likely to be positively or negatively affected by digitization. Second, companies looking to digitize part of their operations (such as company events, product demonstrations, and coding workshops) can use insights from this study to assess the potential benefits of digitizing various aspects of their business. Third, companies could use these results to segment customers according to their preferences for different types of events. Conversely, by understanding which types of events are most likely to benefit from digitization, companies can potentially target specific customer segments more effectively. Finally, the unique, detailed panel data collected from Meetup provide a benchmark for future research in this area, allowing for the generalization of results beyond the scope of this study.

The remainder of the paper is organized as follows. Section 2 describes the data and presents descriptive empirical analyses. In Section 3 we describe the methods, and provide details on the identification of the effects. In Section 4 we assess the heterogeneous impact of digitization on event participation decisions and provide an overview of the robustness checks. Section 5 concludes.

2 Data and Descriptive Analysis

To estimate the heterogeneous impact of event digitization on community participation, we use data from Meetup, a leading global platform for organizing shared-interest groups and events. In this section, we first provide an overview of the Meetup platform. We then describe the data used to estimate the effects of digitization on event participation. We conclude with descriptive evidence that the data provide sufficient variation for estimation.

2.1 Overview of the Meetup Platform

Meetup is an online platform for community building, active since 2002. Meetup's primary purpose is to help shared-interest *groups* organize *events* for their *members*. As of 2020, Meetup provided a platform to 230,000 groups and 50 million members in 193 countries. Many companies use Meetup to develop and support their brand communities, and utilize Meetup's paid service ("Meetup Pro") for managing commercially-oriented groups.

Meetup groups span a wide range of interests, including technology, business, sports, education, and entertainment. When signing up, each group is required to select one of 33 interest

categories that best describes its focal interest. These categories include "Arts and Culture," "Career and Business," "Health and Wellbeing," and "Social Activities."

Group organizers use Meetup's platform primarily to schedule and communicate with members about events. Across groups and interest categories, events can vary widely in their purpose and format, and include workshops, product previews, coding tutorials, conferences, parties, dancing lessons, and book clubs. Most events are in-person, but many are digitized in the form of webinars, virtual conferences and discussion panels, online classes, and asynchronous video feeds. Organizers typically provide members with detailed information on the event format and logistics in an event description. For any group, we expect average member participation might differ between events that are digitized and those that are not. Furthermore, given the variety in the types of groups organizing events through Meetup, we expect the effects of digitizing events could be heterogeneous across groups and interest categories.

Meetup users utilize the platform to discover groups whose interests match their own, and to stay informed about events hosted by those groups. When a *user* of the Meetup platform joins a group, they become a *member* of that group. The main purpose of group membership is to facilitate event planning, as group members are notified about upcoming events, and encouraged to *RSVP*.

The RSVP system allows members to indicate their intention to attend or not attend upcoming events. Both group organizers and the Meetup platform strongly encourage RSVPing for effective event management. However, RSVPing is not mandatory, which may result in fewer RSVPs than the total number of group members. This study focuses on RSVPs as a form of community engagement, and a means of expressing intentions to participate in community events. This is in line with recent work exploring the impact of digitization on intention outcomes, including intention to purchase (Atasoy and Morewedge 2018, Luangrath et al. 2022, Touré-Tillery and Wang 2022, Wiegand and Imschloss 2021), intention to enrol in academic programs (Lee et al. 2022), and intention to donate to charities (Touré-Tillery and Wang 2022). In addition to their practical relevance on Meetup, RSVPs are generally recognized by community and event managers as crucial to estimating head counts, to increasing event awareness among potential attendees, and as a tool to boost prospective attendance (Carpenter 2023, Eventbrite 2023, Tang 2023). Consequently, both the presence and value of event RSVPs are primary outcomes of interest for this study[1].

[1] A complementary measure of members' participation in community events could be members' actual attendance to the events. However, on Meetup, event attendance is not automatically recorded, as group organizers must manually record event attendance for each member (Meetup.com 2018). In practice, attendance records are updated irregularly and inconsistently, and there may be unobserved determinants for updating attendance records that vary by group and event. Therefore, we choose not to use attendance as a measure of community participation. We discuss the plausibility of an intention–behavior gap in this empirical context, as well as

In June 2019, we collected data describing Meetup's groups, events, and members via Meetup's public API. The data cover all publicly available Meetup groups active between January and June, 2019 (24 weeks), in the 15 most populated cities in the U.S. (U.S. Census Bureau 2010). For each group matching these criteria, we collected complete-case information about the group's members and events. The data are divided into two sets. The first set, which comprises all events taking place between January 1 and March 21, 2019 (10 weeks), is used to characterize groups' histories of organizing digitized events and members' past attendance. The remaining 14 weeks of data are used for estimation. Next, we describe the group, event, and member data in detail.

2.2 Data Description

We index each group in the data by g. We seek to characterize heterogeneity in the effects of digitization in terms of groups and interest categories, thus we denote by C_g the interest category for group g. The data also contain a free-text description of the group, as well as indicators for whether new members require approval to join, and whether the organizer uses the Meetup Pro service. We use X_g to collectively refer to all group-level data, apart from the interest category.

For each group, the data contain a list of past and scheduled events. We index events by e, noting that each event is organized by a single group, $g[e]$. The data describing each event include the date and time of the event, a timestamp of when it was first listed on Meetup, a free-text description of the event, and logistical details, such as the location, any entry fee, whether the event is part of a recurring series, and any limits on the number of attendees. We refer collectively to this event-level data as X_e.

Importantly, the data lack indicators for whether an event is digitized. However, the event text descriptions store useful information about the event formats. In Section 2.4 we describe a supervised machine learning procedure for measuring event digitization on the basis of these text descriptions. That procedure yields a continuous treatment variable, $\tilde{D}_e \equiv \widehat{\Pr}[\text{Digitized event}|\text{Event description}]$, representing the expected *degree of digitization* for event e.

The data also contain a list of members for each group g. We index users of the Meetup platform by i, but note that membership always exists in the context of a particular group, g. Hence, a user who is a member of multiple groups may have different member-level characteristics for each group they are a part of. We differentiate between users and members in the text

potential implications of the gap, in Section 5.

when the distinction is relevant.

For this study, the most important information about members is the collection of RSVPs, describing member i's intention to attend event e organized by group g. We denote member i's RSVP for event e by Y_{ie}. Y_{ie} is also the outcome of interest for this study. Observed RSVPs are binary, with $Y_{ie} = 1$ indicating an intention to attend the event, and $Y_{ie} = 0$ indicating an intention not to attend. As noted above, recording an RSVP is not mandatory, leading to missing values for some member-event pairs. We use $Y_{ie} = -1$ to denote a missing RSVP, as we are interested in the impact of event digitization on both presence and value of event RSVPs.

Additional variables in the data help us account for individual heterogeneity in the likelihood of RSVPing to events. In particular, we use the timing of when users joined the platform and became members of different groups, as well as information about events and RSVPs from the first 10 weeks of data, to characterize the typical behavior of members within their groups. From this, we calculate member's participation rate in past group events, and the total number of past group events they have been exposed to on Meetup. We also use positive RSVPs from the first 10 weeks to define metrics related to average event co-attendance among group members. These metrics are the total number of unique peers who also responded positively to the same events, and the share of peers who responded positively to common group events (even if the focal group member did not RSVP positively). These variables, denoted collectively as X_{ig}, are constant across all group events for each member. We also calculate variables that vary by event and member at the point in time when the member creates an RSVP. These are the member's tenure in the group, and the amount of time the event has been posted on Meetup. We refer collectively to these variables as X_{ie}.

In sum, we seek to measure the effect of event digitization (D_e) on members' intentions to participate (Y_{ie}), which we expect to be heterogeneous across groups belonging to different interest categories (C_g). Because event digitization is not random, we include a large number of controls in the estimation procedure, including information about groups (X_g), events (X_e), and members (X_{ig} and X_{ie}). We discuss the identifying assumptions leading to this choice of controls in Section 3.1.

2.3 Estimation Sample

As mentioned, the first 10 weeks of data are used to calculate control variables for member's past exposure to events and average participation, and groups' past digitization of events. The remaining 14 weeks of data are used for estimation, and are structured in a panel format organized by group-event-member-RSVP.

We exclude missing RSVP records for group members who were likely unaware of the existence of an event (6.63% of the data). The awareness status of a member is based on the timestamp of their last visit to their Meetup groups, and on event characteristics that could influence their ability to learn about upcoming events. Appendix A elaborates on the awareness indicator in more detail. The resulting panel, which we refer to as the "full sample," contains 7,851,101 RSVP records, spanning 14 weeks, 15 metro areas (including 508 cities and municipalities), and 33 interest categories. The panel comprises 285,730 members, 118,326 events, and 12,132 groups.

To ensure computational tractability, we estimate the effect of interest on a random subset of the full sample, which we refer to as the "estimation sample". We draw 4,000 random group identifiers from the full sample, and then filter the full sample to only include group, event, and member information for those 4,000 groups. This estimation sample comprises 39,047 events (33% of the total event records), 116,425 members (41% of the total), and 2,684,759 RSVP records (34% of the total), spanning the same weeks, metro areas, and interest categories as the full sample.

2.4 Event Digitization

The unstructured text description for each event stores information on whether the event is digitized. This information allows us to measure event digitization even in absence of a digitization variable in the data. To extract this information, we trained two support vector machines (SVMs) on the event text descriptions. The first predicts whether events have a "digitized" (versus "not digitized") format. As a robustness check, we trained a second SVM to predict events as "in-person" (versus "not in-person"). To train both SVMs, two raters labeled 3025 events from the full sample as "digitized," "in-person," or "both." Disagreements in classification were resolved by a third rater not involved in the initial labeling task. We use the first SVM to predict digitization for all unlabeled events in the data. With 10-fold cross-validation, the SVM achieves 96% prediction accuracy (99% for the second). Appendix B provides further details about the measurement process, as well as descriptive statistics for the predicted cases.

To describe digitization, we use the continuous probability that an event is digitized from the SVM predicting the "digitized" label. This digitization variable ranges between 0 and 1, and represents the expected accuracy of the predicted "digitized" label, $\tilde{D}_e \equiv \widehat{\Pr}[\text{Digitized event} \big| \text{Event description}]$. For example, an event with digitization probability close to 0 indicates that the SVM is very confident that the event has completely non-digital format (Appendix B shows that such an event would also have a high predicted probability of being in-person, based

on predictions from the second SVM). The majority of events are labelled consistently across the two models, indicating either high in-person probability and low digitization probability, or vice versa. A small subset of events (0.4%) has different labels predicted by the two SVMs. This suggests they might incorporate both in-person and digital components.

The use of a continuous measure of event digitization (as opposed to categorical or binary) has several advantages. First, a continuous variable offers a finer-grained representation of the likelihood of event digitization. This granularity is helpful in the analysis of subtle variations in the degree of digitization for different events. Second, in these real-world events, the demarcation between digitized, in-person, and hybrid events may not be clear-cut. Some events may feature both digitized and in-person elements, with different combinations of elements catering to different community needs, and no quantitative rule to map different combinations of elements into categorical levels. On the other hand, intermediate digitization probabilities could better characterize different combinations of digitized and in-person elements, and would not require additional assumptions. In the robustness checks (Section 4.3), we allow the event digitization variable to resemble a binary treatment, and categorize events as "digitized" ($\tilde{D}_e > 50\%$) versus "non-digitized" ($\tilde{D}_e \leq 50\%$).

Finally, we use the digitization variable to measure the group-varying average rate of past event digitization at the time of creating each event entry, as it helps to differentiate among groups that differ in their baseline propensity to digitize events.

2.5 RSVPs

Recall that we use the variable Y_{ie} to represent member i's RSVP for event e, and treat it as a measure of participation. Also recall that while the Meetup platform and group organizers strongly encourage members to RSVP, creating RSVPs is not mandatory, which leads to missing responses.

We note that not all missing RSVP entries in the panel are due to a conscious decision not to respond. Some RSVPs may be missing because members were unaware of particular events (e.g., because they did not visit the group's page or the platform during the period between when the event was posted and when it took place). For such members, the absence of an RSVP is not due to deliberation, but it is the only available option — thus unrelated to event digitization. To identify members who are potentially unaware of particular events, we define an indicator of event awareness for each member-event pair. This variable is derived from event and member information, including the time of their last visit to the platform (see Appendix A for details). Using this variable, as mentioned previously, we exclude 6.63% of member-event

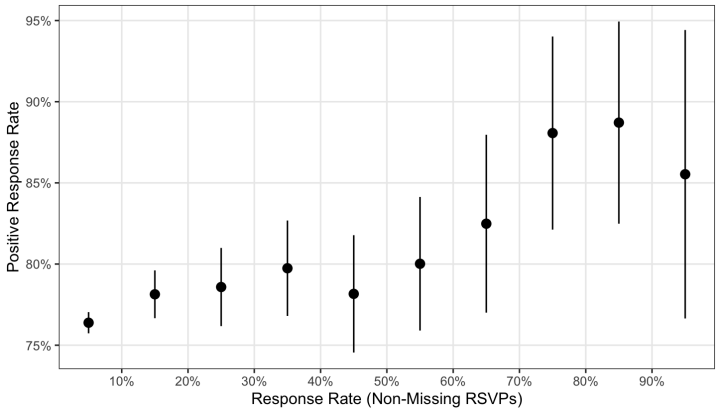


Figure 1: Less Attractive Events Have Lower Response Rates. *Note:* Response rates for 18,566 events (10% of the full sample) are shown. Error bars represent the 95% Gaussian confidence limits of the proportions.

observations where the member was likely unaware of the event, and thus did not RSVP.

After excluding non-responses from unaware members, the absence of an RSVP provides information about the attractiveness of digitized events. For example, a member interested in attending an event, but unable to join, might be more likely to RSVP with $Y_{ie} = 0$ than an individual who has no interest in the event. Conversely, a member who has no interest in an event might be less likely to RSVP at all, than an individual who is interested in the event. In short, we surmise that missing RSVPs are not missing at random, and thus could provide information about the attractiveness of events for some members.

We therefore investigate whether the proportion of missing RSVPs correlates with event attractiveness, by plotting data from a random sample of 18,566 events (10% of the full sample; Figure 1). The figure shows that events with greater proportions of negative (versus positive) RSVPs also have greater proportions of non-responses.

Based on this evidence, we make two important assumptions about the RSVP variable. First, because creating a negative RSVP (unlike a non-response) requires effort on the part of a group member, and thus reflects a degree of engagement with the event and/or the group, we assume that the outcome is an *ordered* categorical variable with three levels: non-response $<$ negative response $<$ positive response. Second, we assume that the non-responses in the estimation sample are the product of members' choices and can be modeled in the same way as positive or negative RSVPs. In the Online Appendix, we also examine the relationship between non-responses and positive RSVPs among groups that contribute the most to the identification of the measured effect, specifically groups that organize a mix of digitized and in-person events. This analysis suggests that non-response rates do not depend on event digitization.

	N. Events			N. Members Exposed to Format		
	Digitized/Not In-Person	In-Person/Not Digitized	Both Formats	Digitized/Not In-Person	In-Person/Not Digitized	Both Formats
Model	(1)	(2)	(3)	(4)	(5)	(6)
SVM Digitization	608	38336	103	355	105868	10202
SVM In-Person	621	38346	80	416	105422	10587

Table 1: The Estimation Sample Provides Variation in the Extent of Event Digitization Within the Same Group. *NOTE:* The number of events classified by format (columns 1-3), and the number of members exposed to different combinations of event formats (columns 4-6) are shown. The classification into "digitized", "in-person", and "both" event class was performed using a 50% threshold for the predicted SVM probabilities ("digitized" if PR[Online] $> 50\%$; "in-person" if PR[Online] $< 50\%$, and "both" if PR[Online] $= 50\%$). The numbers refer to the estimation sample. Appendix Table 10 reports the same statistics for the full sample.

2.6 Exposure to Digitized Events

To estimate the causal effect of digitization on member participation, it is important that there be variation in event digitization within the same group. This variation allows the same members to be exposed to events with different features while keeping the member and group characteristics constant. Table 1 summarises the extent of this variation, and presents the number of events by digitization format, as well as the number of members exposed to each format. Based on the SVM predictions, 154 groups in the estimation sample (3.8% of the groups in the sample) organized both digitized and non-digitized events; 31 groups (0.8%) organized exclusively digitized events; and 3815 groups (95.4%) organized exclusively in-person events.

Based on the same SVM predictions, 608 events in the estimation sample were digitized (col. 1, Table 1, 1.5% of the events in the sample); 38336 events were in-person (col. 2, 98.2%); and the format of 103 events contained both digitized and non-digitized elements (col. 3, 0.3%).

This variation means Meetup members were exposed to events with different degrees of digitization, both within and across groups. Most Meetup members were exposed only to in-person/non-digitized formats (col. 5, Table 1, 90.9% of the sample), consistent with a setting where in-person events were the norm. The second largest group, however, comprises members who were exposed to both formats during the observation period (col. 6). This group accounts for 8.8% of members in the estimation sample, and provides important variation to statistically identify the effects of interest. Finally, a small percentage of members were exposed only to digitized/non-in-person events (col. 4, 0.3%).

2.7 Group Categories

An organizer's decision to digitize an event may be influenced by several factors, such as cost-effectiveness, organizational flexibility, and the need for networking opportunities (6Connect

	Event Digitization Probability											
Category	0%	1–10%	11–20%	21–30%	31–40%	41–50%	51–60%	61–70%	71–80%	81–90%	91–100%	Total
Arts Culture	0.79	0.15	0.04	0.01	0.00	0.00	0.01	0.00	0.00	0.00	0.00	98
Book Clubs	0.78	0.20	0.01	0	0	0	0	0	0	0	0	69
Career Business	0.60	0.19	0.06	0.03	0.02	0.01	0.01	0.01	0.01	0.01	0.04	785
Cars Motorcycles	0.91	0.10	0	0	0	0	0	0	0	0	0	21
Community Environment	0.58	0.22	0.08	0.03	0.03	0.01	0.01	0.02	0.01	0.01	0	108
Dancing	0.84	0.15	0.01	0	0	0	0	0	0	0	0	113
Education Learning	0.73	0.19	0.03	0.02	0.01	0	0	0	0.01	0.01	0.01	136
Fashion Beauty	1	0	0	0	0	0	0	0	0	0	0	14
Fitness	0.80	0.13	0.05	0	0.01	0	0	0.01	0.01	0	0	103
Food Drink	0.87	0.11	0.02	0	0	0	0	0	0	0	0	123
Games	0.80	0.14	0.02	0.01	0	0.02	0	0	0	0	0	142
Government Politics	0.67	0.28	0.02	0	0	0.02	0	0	0	0	0	43
Health Wellbeing	0.63	0.21	0.05	0.02	0.01	0.02	0.01	0.02	0.01	0.01	0.02	378
Hobbies Crafts	0.88	0.06	0.04	0.02	0	0	0	0	0	0	0	49
Language	0.82	0.13	0.02	0.01	0	0.01	0.01	0	0	0	0.01	181
Lgbt	0.72	0.19	0.04	0.01	0.02	0	0	0.01	0	0	0	83
Lifestyle	1	0	0	0	0	0	0	0	0	0	0	5
Movies Film	0.73	0.16	0.05	0.01	0.01	0.01	0.03	0	0	0	0	96
Music	0.82	0.15	0.03	0.01	0	0	0	0	0	0	0	115
New Age Spirituality	0.64	0.23	0.03	0.03	0.01	0.01	0.03	0	0.01	0.01	0.01	319
Outdoors Adventure	0.76	0.15	0.03	0.03	0	0	0.01	0	0	0	0.01	267
Paranormal	0.57	0.14	0.14	0.14	0	0	0	0	0	0	0	7
Parents Family	0.81	0.11	0.04	0	0	0	0.04	0	0	0	0	27
Pets Animals	0.91	0.07	0.02	0	0	0	0	0	0	0	0	46
Photography	0.65	0.20	0.07	0.02	0.01	0.01	0.01	0.02	0	0.01	0	91
Religion Beliefs	0.74	0.20	0.01	0.02	0.01	0	0.01	0.01	0	0	0	105
Sci–Fi	0.77	0.23	0	0	0	0	0	0	0	0	0	35
Singles	0.71	0.16	0.06	0.02	0	0	0.02	0	0	0.02	0	87
Socializing	0.74	0.15	0.04	0.01	0.01	0.01	0.01	0	0.01	0.01	0.01	391
Sports Recreation	0.88	0.08	0.01	0.01	0.01	0	0.01	0	0	0	0	170
Support	0.73	0.20	0.02	0.02	0	0	0.02	0	0	0	0	44
Tech	0.64	0.18	0.05	0.02	0.01	0.01	0.01	0.01	0.02	0.01	0.03	1019
Writing	0.68	0.19	0.04	0.04	0.02	0.01	0	0	0	0.01	0.01	84

Table 2: The Proportion of Digitized Events Varies across Interest Categories. *Note:* The proportions of events by interest category and predicted probability of digitization refer to the estimation sample.

2020). Hence, events in certain interest categories, such as business and career development, may be more likely to be digitized than events in other categories, such as socialization and social support. As shown in Table 2, some categories, such as "Fashion and Beauty," "Pets and Animals," or "Cars and Motorcycles," have high proportions of in-person events; whereas groups in categories like "Technology," "Community and Environment," and "Career and Business" have high proportions of digitized events.

Table 2 also suggests a sparsity problem in the estimation sample, with many of the interest categories having low proportions of digitized events. We seek to simultaneously reduce the number of group categories and address the sparsity issue by estimating an LDA topic model on the names and descriptions of groups. Using quantitative criteria from Arun et al. (2010), Cao et al. (2009), Deveaud et al. (2014) and Griffiths and Steyvers (2004), we set the number of latent topics at 14 (further details about the LDA procedure are provided in Appendix C).

Table 3 describes the 14 interest topics. The topic labels are qualitative interpretations of shared-interest categories, derived from the words with the highest probability of associating with each topic, as well as the interest categories most associated with each topic. Table 4 shows the proportion of digitization probabilities per topic, revealing that the sparsity problem has been mitigated.

The LDA model does not assign a unique topic to each group, but rather considers each group to be a mixture of topics, with weights given by a vector of topic probabilities. These

Topic	Label	Interpretation
1	Organization and Meta-Events	Organizational meetings, meta-events (i.e. events about event organization), event information sessions
2	Social Support and Causes	Social support and social causes
3	Language and Toastmaster	Languages, toastmaster meetings, foreign language courses
4	Yoga and Wellness	Yoga and meditation, mindfulness, health and well-being, guided meditations
5	Socializing and Matchmaking	Socialization events, food and drinks, matchmaking
6	Investments and Networking	Investments and business networking, real estate, property investments
7	Board and Card Games	Board games, card games, entertainment
8	Arts and Music	Music, arts, dancing, cultural events
9	Spirituality and Healing	Spirituality, health, energy healing, nature, spiritual support
10	Sports and Photography	Sports, sport clubs, photography
11	Outdoor Activities	Outdoor and outdoor activities – running, riding, cycling, nature, camping, walking
12	Business Networking and Careers	Business networking, professional events, career events, startups, industry events
13	Writing and Hobbies	Writing clubs, skill development workshops, religion, hobbies
14	Technology and Software	Technology, software development, coding, design, programming, web

Table 3: Summary of the 14 Latent Topics Obtained from the LDA Procedure.

	Digitization Probability (Estimation Sample)										
Topic	0%	1-10%	11-20%	21-30%	31-40%	41-50%	51-60%	61-70%	71-80%	81-90%	91-100%
1 - Organization	0	0.9540	0.0212	0.0064	0.0087	0.0037	0.0037	0.0009	0	0.0009	0.0005
2 - Social Support	0.0007	0.9508	0.0362	0.0065	0.0014	0.0007	0.0007	0.0007	0	0	0.0022
3 - Languages	0	0.9728	0.0152	0.0060	0.0005	0.0005	0	0	0	0.0016	0.0033
4 - Yoga	0.0005	0.9322	0.0221	0.0019	0.0063	0.0053	0.0048	0.0096	0.0010	0.0111	0.0053
5 - Socializing	0.0003	0.9790	0.0146	0.0017	0.0015	0	0.0007	0.0008	0.0001	0.0007	0.0006
6 - Investments	0	0.8303	0.0526	0.0319	0.0154	0.0047	0.0047	0.0101	0.0024	0.0024	0.0455
7 - Games	0.0003	0.9769	0.0108	0.0066	0	0.0003	0.0039	0	0	0	0.0012
8 - Dance	0.0003	0.9731	0.0172	0.0058	0.0015	0	0.0006	0	0.0003	0	0.0012
9 - Spirituality	0	0.9246	0.0399	0.0120	0.0040	0.0040	0.0049	0.0022	0.0013	0.0049	0.0022
10 - Sports	0.0026	0.9812	0.0115	0.0026	0.0003	0.0003	0	0.0015	0	0	0
11 - Outdoors	0	0.9737	0.0096	0.0085	0.0007	0.0021	0.0010	0.0007	0.0003	0.0010	0.0024
12 - Business	0.0005	0.9006	0.0453	0.0118	0.0118	0.0041	0.0021	0.0051	0.0021	0.0026	0.0139
13 - Writing	0	0.9572	0.0169	0.0054	0.0030	0.0013	0.0054	0.0003	0.0037	0.0051	0.0017
14 - Tech	0	0.8134	0.0496	0.0185	0.0053	0.0029	0.0144	0.0156	0.0168	0.0361	0.0275

Table 4: The LDA Procedure Mitigates the Problem of Sparsity in the Extent of Event Digitization across Interest Categories. *Note:* The proportions of events by interest topic and digitization probability refer to the estimation sample. Each group is assigned to one topic, based on the highest group-topic classification weight. The topics are described in Table 3.

weights provide a measure of how well the group matches each of the topics. When estimating the effect of event digitization on member participation, we use these 14 topic weights in place of the binary categories, and denote them $\tilde{C}_g$.

In sum, the data contain a rich description of members, events, and groups. Table 5 provides summary statistics for the variables in the estimation sample.

3 Estimating the Heterogeneous Effect of Event Digitization

In this section, we describe our strategy for measuring the effect of digitizing events on members' intent to participate. Recall that we denote member i's RSVP to event e as Y_{ie}, and that Y_{ie} can take on one of three values: attending ($Y_{ie} = 1$), not attending ($Y_{ie} = 0$), or missing ($Y_{ie} = -1$). $\tilde{D}_e$ denotes the estimated (continuous) degree of digitization for event e, organized by group $g[e]$. Group $g[e]$ is partially characterized by a vector of LDA topic probabilities, $\tilde{C}_g$. Our goal is to estimate the causal effect of event digitization on RSVP responses for the groups in the data, and to understand how heterogeneity in these effects varies among groups catering to

Description	Mean	SD
Event Digitization ($\tilde{D}_e$)		
Pr[Digitized]	0.04	0.11
RSVPs (Y_{ie})		
Positive RSVPs	0.07	0.26
Negative RSVPs	0.02	0.15
Non-responses	0.90	0.29
Group Topics (C_g, Weights $\in$ [0,1])		
Topic 1 - Organization and Meta-Events	0.05	0.14
Topic 2 - Social Support and Causes	0.06	0.19
Topic 3 - Language and Toastmaster	0.05	0.16
Topic 4 - Yoga and Wellness	0.11	0.24
Topic 5 - Socializing and Matchmaking	0.06	0.18
Topic 6 - Investments and Networking	0.12	0.27
Topic 7 - Board and Card Games	0.06	0.17
Topic 8 - Arts and Music	0.05	0.16
Topic 9 - Spirituality and Healing	0.04	0.15
Topic 10- Sports and Photography	0.13	0.26
Topic 11- Outdoor Activities	0.06	0.19
Topic 12- Business Networking and Careers	0.05	0.18
Topic 13- Writing and Hobbies	0.09	0.24
Topic 14- Technology and Software	0.07	0.20
Other Group Features (X_g)		
Members (×1000)	0.15	0.25
Is Open {0,1}	0.91	0.29
Is Pro {0,1}	0.04	0.20
Event Features (X_e)		
Description Length (×10000 characters)	0.14	0.13
Has Fee {0,1}	0.02	0.14
Has Limits {0,1}	0.30	0.46
Has Venue {0,1}	0.91	0.28
Is Series {0,1}	0.05	0.23
Morning {0,1}	0.22	0.42
Avg. Digitization in Group	0.04	0.11
Member Features Varying by Event (X_{ie})		
Tenure at Time of RSVP (×10 years)	0.45	0.36
Time Since Event Creation (×10 years)	0.03	0.08
Member Features Varying by Group (X_{ig})		
Avg. Past Response in Group	0.33	0.34
N. Co-Attending Peers (log1p)	0.77	0.59
Share of Co-Attending Peers (%)	0.16	0.10
N. Past Events in Group (×100)	0.16	0.33

Table 5: Summary Statistics for the Estimation Data

different interest topics (i.e., with different values of $\tilde{C}_g$).

3.1 Identification Strategy

Estimating the effect of event digitization on members' responses is challenging due to potential confounding between the organizer's decision to digitize an event and the members' RSVP responses. Both can be jointly influenced by several factors, such as varying demand for digitization across groups, location, and time. To address these potential sources of bias, we include

a rich set of control variables at the member, event, and group levels, collated and denoted as

$$W_{ie} \equiv \begin{pmatrix} X_e & X_{ig[e]} & X_{ie} \end{pmatrix} \quad \text{and} \tag{1}$$

$$W_g \equiv \begin{pmatrix} X_g & \tilde{C}_g \end{pmatrix}, \tag{2}$$

where, as noted previously, $\tilde{C}_g$ and X_g contain variables that vary by group, X_e contains variables varying by event, X_{ig} contains variables varying by group member, and X_{ie} contains variables varying by member and event. In addition to W_{ie} and W_g, we control for other unobserved demand shocks using fixed effects that vary by the day $t[e]$ and the location $m[e]$ of event e, denoted τ_t and ζ_m respectively. We further control for any remaining group-specific influences, either parametrically in a hierarchical model (η_g) or non-parametrically through clustered standard errors.

To identify the causal effect of event digitization on RSVPs, we make three behavioral assumptions. First, we assume that member i's unobserved demand for event e, which we denote ϵ_{ie}, is conditionally independent of event digitization given W_{ie}, W_g, and the fixed effects:

$$\epsilon_{ie} \perp\!\!\!\perp D_e \big| W_{ie}, W_g, \tau_t, \zeta_m, \eta_g \tag{A1}$$

Under (A1), no individual in a Meetup group can be influential enough that their demand for an event can influence the decision to digitize it. Furthermore, (A1) implies that, conditional on W_{ie}, W_g, and the fixed effects, any remaining correlation between the individual demand for an event and the decision to digitize the event is negligible.

Second, we assume that the unobserved demand from any Meetup member is conditionally independent of average demand among the other group members, $\bar{\epsilon}_{i'e}$, given D_e, W_{ie}, W_g, and the fixed effects.

$$\epsilon_{ie} \perp\!\!\!\perp \bar{\epsilon}_{i'e} \big| D_e, W_{ie}, W_g, \tau_t, \zeta_m, \eta_g \tag{A2}$$

Under (A2), no Meetup member's demand for an event can be so influential as to shift average demand among all other group members.

The set of control variables and identifying assumptions discussed so far imply that the treatment assignment (event digitization) is as good as random for a Meetup member conditional on the observed data, and under the identifying assumptions described above (Rosenbaum and Rubin 1983). Figure 2 depicts the main non-parametric identifying assumptions in DAG format.

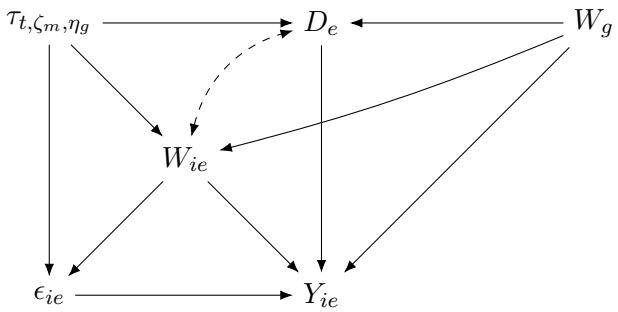


Figure 2: Directed Acyclic Graph.

3.2 Structural Causal Model

We recover the estimated effect of D_e on Y_{ie} with heterogeneous parameter estimates in a structural causal model (SCM) that reflects the differential impact of digitization across groups belonging to different interest topics. These parameters are identified from the data, as members are exposed to events with different levels of digitization within and across groups. The statistical identification arguments for the effects of other event-level and group-level parameters on RSVP choices are similar to those for standard choice models using panel data, where the same member's RSVP decisions are observed over multiple choice occasions. Event-level parameters are identified from members making RSVP choices about events with varying characteristics organized by the same group over time, and similar events organized by different groups at the same time. Group-level parameters are identified from the same members making RSVP choices for different events.

The SCM is a hierarchical, discrete choice model, which we estimate in a Bayesian framework. Each member i chooses whether and how to respond to event e organized by group g, with Y_{ie} representing their RSVP value, including the option not to response. As described in Section 2.5, we do not assume that RSVPs are missing at random, but rather model Y_{ie} with a censoring parameter in the model, L, which delineates between negative RSVPs recording that the member will not attend, and non-responses indicating the member will not attend.

Choice Model. We define u_{ie} as the utility that member i expects from attending an upcoming event e. This event is organized at time t, by group g, in market m. Based on the discussion in Sections 2.6 and 3.1, we anticipate the expected utility from attending this event could be influenced by factors that vary by market (m), time (t), group (g), member (i), and event (e). In particular, the model allows the extent of estimated event digitization, $\tilde{D}_e$, to have a direct, but possibly heterogeneous impact on the utility from attending the event. Specifically, the model allows the effect of event digitization to differ across groups, as determined by their

LDA loadings on 14 interest topics ($\tilde{C}_g$). These determinants of member i's RSVP decision are reflected in the following expected utility function:

$$u_{ie} = \tilde{D}_e \cdot \tilde{C}_g' \delta + W_{ie}' \beta + \tau_t + \zeta_m + \eta_g + \epsilon_{ie}, \tag{3}$$

where the vector δ represents the heterogeneous effects of digitization; τ_t, ζ_m, and η_g are fixed effects for market (city), time, and group; W_{ie} represents control variables that co-vary by individual and event/group; and ϵ_{ie} represents unobserved, idiosyncratic factors affecting i's expected utility from event e.

As discussed in Section 2.5, we jointly model individual i's decision to record a positive or negative RSVP, or no response at all. We assume these three RSVP values are ordered in terms of their utility, such that censoring occurs only among individuals who are not planning to attend. This specification requires two thresholds. The first, which we set to 0, separates positive RSVPs ($Y = 1$) from negative RSVPs ($Y = 0$). The second, which we set to $L < 0$, separates non-responses ($Y = -1$) from explicitly recorded, negative RSVPs. Equation 4 describes the relationship between utilities (u_{ie}), the thresholds (0 and L), and the observed outcomes (Y_{ie}):

$$Y_{ie} = \begin{cases} 1 & \text{if } u_{ie} > 0 \\ 0 & \text{if } L < u_{ie} \leq 0 \\ -1 & \text{if } u_{ie} \leq L \end{cases} \tag{4}$$

If the utility from the event is greater than zero ($u_{ie} > 0$), individual i will leave a positive RSVP. Otherwise, the outcome is determined by the censoring parameter L. When $u_{ie} \leq L$, the individual does not leave an RSVP. Otherwise, when $L < u_{ie} \leq 0$, the individual leaves a negative RSVP, indicating they will not attend. Assuming that ϵ_{ie} follows a standard normal distribution, and denoting $v_{ie} = u_{ie} - \epsilon_{ie}$, the model implies the following ordered probit likelihood:

$$\mathcal{L}_{ie} = \ell(Y_{ie}) \tag{5}$$

$$\ell(1) = \Pr[v_{ie} + \epsilon_{ie} > 0] = 1 - \Phi(-v_{ie}) \tag{6}$$

$$\ell(0) = \Pr[L < v_{ie} + \epsilon_{ie} \leq 0] = \Phi(-v_{ie}) - \Phi(L - v_{ie}) \tag{7}$$

$$\ell(-1) = \Pr[v_{ie} + \epsilon_{ie} \leq L] = \Phi(L - v_{ie}) \tag{8}$$

where Φ is the CDF of the standard normal distribution. Finally, we derive a Bayesian posterior

distribution over the model parameters by specifying the following prior distributions:

$$\eta_g \sim N\left(W_g'\gamma, 1\right) \tag{9}$$

$$\delta, \beta, \gamma, \zeta_m, \tau_e \overset{\text{iid}}{\sim} N(0,1) \tag{10}$$

$$L \sim N^{-}(0,1) \tag{11}$$

where N^{-} indicates a standard normal distribution truncated above at 0. During the estimation, we normalize the first element of the fixed-effect vectors τ_t and ζ_m to 0. The causal effect of interest is captured by the parameter estimates for the vector δ.

Parameter Estimation and Model Validation. We obtain the posterior parameter estimates using the shell interface to Stan (Stan Development Team 2021). We use a No-U-Turn variant of the Hamiltonian Monte Carlo algorithm to sample from the posterior distribution of the model parameters. Finally, we validate the Bayesian estimation procedure using simulation-based calibration (SBC; Talts et al. 2018).

Counterfactual Policy Evaluation. We use the parameter estimates from the SCM to evaluate the impact of shifting all digitized events from fully in-person to fully online. To construct the counterfactual estimation sample, we first extract the subset of events for which (i) the probability of digitization (from the first SVM) is greater than 50%, (ii) the probability of being in-person (from the second SVM) is less than 50%, and (iii) the corresponding group had organized at least one in-person event in the past. We base this counterfactual analysis on digitized events for two reasons. One is that digital events can typically be made in-person, whereas converting in-person events to be fully digital may not always be feasible. The other is that members of groups organizing digital events are presumed to have some experience with digitized formats, where again the reverse may not always be true.

For each event in the counterfactual subset, we simulate members' responses, first with $\tilde{D}_e$ set to 0 (non-digitized) and then with $\tilde{D}_e$ set to 1 (fully digitized), using the full posterior distribution of the structural parameters. We then compare the distributions of positive, negative, and missing RSVPs across the two simulated scenarios in the counterfactual policy to obtain the average treatment effect on the treated (ATT). To better quantify the heterogeneity in the effect of digitization on community participation, we also estimate conditional ATTs (CATTs) by group and by interest topic. In the latter case, we assign groups to the interest topic with the highest probability, as estimated by the LDA model.

3.3 Robustness Checks: Causal Random Forests and Nearest-Neighbor Matching

We check the robustness of the structural estimates in two ways. First, we estimate causal random forests (CRFs; Wager and Athey 2018). CRFs relax functional form assumptions on the structure of the unobserved errors, as well as on the distribution of group-level effects, while achieving many of the desirable statistical properties of regression-based methods, such as asymptotic consistency. The CRF regressions are based on the same set of causal assumptions described in Section 3.1.

In implementing the CRFs, we assume that there is considerable heterogeneity across groups, and that there could be unobserved group-level features that serve as treatment effect moderators (e.g. strength of group leadership, susceptibility to social influence, or group norms). For computational reasons, including 4,000 group-level fixed effects is not feasible. Hence, we perform a group-level cluster-robust analysis as an alternative. We estimate an average treatment effect of digitization using a binarized treatment indicator, $\check{D}_e$, taking a value of 1 when the probability of event digitization exceeds 50%, and 0 otherwise. The binarized treatment indicator provides an alternative to the assumption of a continuous treatment variable, and allows us to estimate average treatment effects on treated units using the *grf* package in R (Tibshirani et al. 2021).

With this set-up, we aim at estimating the following conditional average treatment effect on the treated, heterogeneous by interest topic:

$$CATT_c = E\left[\frac{Cov[\check{D}_e, Y_{ie}|W_{ie}, W_g, t, m]}{Var[\check{D}_e|W_{ie}, W_g, t, m]}\middle|\check{D}_e = 1, \check{C}_g = c\right] \tag{12}$$

where $\check{C}_g$ is the interest category with the greatest probability for group g.

In addition to the CRFs, we perform and evaluate the results of a nearest-neighbor matching procedure on the identifying pre-treatment covariates (Ho et al. 2011). With matching, we address potential concerns about covariate imbalance between digitized and in-person events. We perform nearest neighbor matching with replacement, on propensity scores estimated using a logistic regression of the binary digitization treatment ($\check{D}_e$) on the identifying covariates, W_{ie} and W_g. After matching, the distributions of covariates in digitized and in-person events are approximately similar, and close to what they would be in a randomized experiment (Appendix D). Using the matched data, we then calculate CATTs of digitization by group.

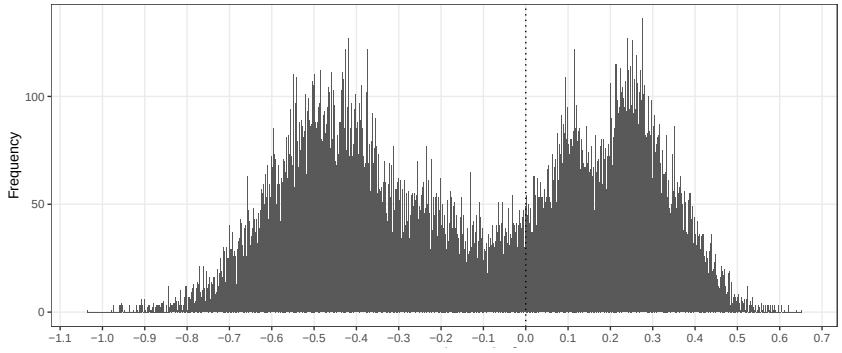


Figure 3: The Average Effect of Digitization on the Utility from RSVPing is Negative but Highly Heterogeneous. *Note:* Posterior density of δ shown on the utility scale. N = 56000, Mean = -0.13, Std. Deviation = 0.35, Median = -0.14, Min. = -1.04, Max. = 0.65. Hartigans' dip test for unimodality (H_0: distribution is unimodal): D = 0.042, p-value < 0.001.

4 Results

This section presents the results of the models detailed in Section 3. Section 4.1 discusses the key parameter estimates for the heterogeneous effect of digitization on event participation, across different interest topics. Section 4.2 evaluates the heterogeneous effects of a counterfactual policy that shifts all digitized events from fully in-person to fully online. In Section 4.3, we assess the robustness of the results to alternative model specifications.

4.1 Structural Causal Model

The MCMC sampler's diagnostics indicate that the posterior distributions of the model parameters are sampled successfully. Specifically, the NUT sampler does not report divergences, which indicates the Hamiltonian Markov Chain sufficiently explores the target distribution in Equation (8). Moreover, the SBC validation provides satisfactory evidence of the model's ability to recover the parameter estimates accurately. Additional detail on model diagnostics is available in the Online Appendix.

Table 6 summarizes the quantiles of the posterior distributions for the parameters in Equation 3. The results indicate substantial heterogeneity in the impact of digitization across interest topics (parameter vector δ), ranging from very negative (e.g. $\hat{\delta}^{\text{Social Support}} \equiv \mathbb{E}[\delta^{\text{Social Support}}|\text{data}] = -0.57$), to very positive (e.g. $\hat{\delta}^{\text{Games}}$ and $\hat{\delta}^{\text{Tech}} = 0.33$) on the utility scale.

Figure 3 presents only the posterior distribution of the parameter vector δ. The distribution implies an overall negative impact of digitization on event participation (Mean $= -0.105$; 95% weighted C.I. $= [-0.113; -0.096]$) when averaged over events in the estimation sample. However, the posterior distribution of δ also reflects the highly heterogeneous impact of digitization on members' participation utility across events in various interest topics.

Variable	Mean	SE	SD	2.5% CI	50% CI	97.5% CI	ESS	$\hat{R}$
Topics × Digitization (δ)								
Organization and Meta-Events × Digital	0.240	0.001	0.040	0.170	0.240	0.300	2056	1
Social Support and Causes × Digital	-0.390	0.002	0.090	-0.540	-0.390	-0.240	2946	1
Language and Toastmaster × Digital	-0.480	0.001	0.084	-0.620	-0.480	-0.340	3694	1
Yoga and Wellness × Digital	0.032	0.001	0.079	-0.100	0.031	0.160	5164	1
Socializing and Matchmaking × Digital	-0.210	0.001	0.075	-0.340	-0.210	-0.088	3360	1
Investments and Networking × Digital	0.110	0.001	0.043	0.039	0.110	0.180	2211	1
Board and Card Games × Digital	0.330	0.002	0.099	0.170	0.330	0.490	4546	1
Arts and Music × Digital	-0.380	0.002	0.110	-0.570	-0.380	-0.190	4079	1
Spirituality and Healing × Digital	-0.570	0.002	0.110	-0.750	-0.570	-0.380	5812	1
Sports and Photography × Digital	-0.530	0.002	0.120	-0.730	-0.520	-0.340	5827	1
Outdoor Activities × Digital	-0.520	0.002	0.140	-0.760	-0.520	-0.290	3968	1
Business Networking and Careers × Digital	0.240	0.001	0.075	0.110	0.240	0.360	3449	1
Writing and Hobbies × Digital	-0.026	0.002	0.120	-0.230	-0.026	0.170	4390	1
Technology and Software × Digital	0.330	0.001	0.070	0.220	0.330	0.440	5234	1
Group Characteristics (γ_g)								
Organization and Meta-Events	-0.430	0.008	0.096	-0.580	-0.430	-0.250	132	1
Social Support and Causes	-0.440	0.007	0.093	-0.590	-0.440	-0.280	174	1
Language and Toastmaster	-1.000	0.008	0.097	-1.100	-1.000	-0.820	148	1
Yoga and Wellness	-1.100	0.013	0.130	-1.400	-1.200	-0.940	100	1
Socializing and Matchmaking	-1.100	0.008	0.100	-1.300	-1.100	-0.930	173	1
Investments and Networking	-0.800	0.009	0.099	-0.970	-0.800	-0.650	127	1
Board and Card Games	-1.100	0.010	0.110	-1.300	-1.100	-0.900	127	1
Arts and Music	-1.000	0.007	0.090	-1.100	-1.000	-0.840	151	1
Spirituality and Healing	-1.000	0.009	0.110	-1.200	-1.000	-0.820	152	1
Sports and Photography	-1.200	0.010	0.110	-1.400	-1.200	-1.000	135	1
Outdoor Activities	-0.640	0.008	0.110	-0.820	-0.640	-0.470	168	1
Business Networking and Careers	-1.300	0.008	0.120	-1.500	-1.300	-1.100	210	1
Writing and Hobbies	-0.950	0.008	0.110	-1.100	-0.960	-0.750	198	1
Technology and Software	-0.960	0.009	0.110	-1.100	-0.970	-0.770	141	1
Group Size	-0.750	0.009	0.067	-0.860	-0.760	-0.640	57	1
Open Group	-0.047	0.004	0.056	-0.140	-0.049	0.046	181	1
Pro License	-0.130	0.005	0.072	-0.240	-0.120	-0.006	199	1
Event, Member-Event, and Member-Group Characteristics (β)								
Avg. Digitization in Group	-0.280	0.001	0.038	-0.340	-0.280	-0.220	1697	1
Cap on RSVPs	0.069	0.000	0.005	0.062	0.069	0.076	4398	1
Event Description Length	0.600	0.000	0.017	0.570	0.600	0.630	3987	1
Event Fee Charged	-0.250	0.000	0.016	-0.280	-0.250	-0.220	3161	1
Morning Event	0.022	0.000	0.004	0.015	0.022	0.029	5854	1
Recurring Event	-0.100	0.000	0.013	-0.120	-0.100	-0.080	4890	1
Venue is Listed	0.034	0.000	0.006	0.024	0.034	0.044	7126	1
Tenure	0.030	0.000	0.004	0.024	0.030	0.036	9452	1
Time of Response	0.280	0.000	0.023	0.250	0.280	0.320	3816	1
Avg. Response Rate	0.860	0.000	0.008	0.840	0.860	0.870	4248	1
N.Co-Attendees	0.660	0.000	0.002	0.660	0.660	0.670	5173	1
Past Events Exposure	-0.110	0.000	0.006	-0.120	-0.110	-0.097	8061	1
Share Co-Attendees	-1.100	0.000	0.017	-1.200	-1.100	-1.100	7086	1

Table 6: Posterior Parameter Estimates (Utility Scale). *Note:* Posterior distributions estimated using a Hamiltonian Monte Carlo algorithm. Posterior statistics calculated over 4 chains, 1000 iterations per chain. Specification estimated: $u_{ie} = D_e \cdot \tilde{C}'_g\delta + W'_{ie}\beta + \tau_t + \zeta_m + \eta_g + \epsilon_{ie}$; $Y_{ie} = 1$ if $u_{ie} > 0 > L$; $Y_{ie}0$ if $L < u_{ie} \leq 0$; $Y_{ie} = -1$ if $u_{ie} \leq L$. Priors: $\eta_g \sim N(X'_g\gamma_g, 1)$; $\delta, \beta_c, \beta_e, \mu, \gamma, \zeta_m, \tau_e \overset{\text{iid}}{\sim} N(0,1)$; $L \sim N^-(0,1)$. Estimated censoring threshold parameter L: mean -0.20, SD 0.00.

Interpreting the average negative impact of digitization without considering the high degree of heterogeneity in these effects may result in misleading managerial implications. Therefore, to further characterize the heterogeneous impact of digitization, Figure 4 shows the posterior distributions of the treatment effect parameters by interest topic (i.e., the individual elements of the vector δ). The results indicate that digitization hinders participation in events related to socializing (dating, clubbing, social drinking, fun and group adventures, city trips; $\hat{\delta}^{Socializing} = -0.21$), social support and social causes (support among women, local politics, environment, local chapters of national organizations; $\hat{\delta}^{SocialSupport} = -0.39$), sports, sport clubs and photography (recreational sports, photography competitions, tennis and soccer, local clubs; $\hat{\delta}^{Sports} = -0.53$), languages and public speaking (toastmaster meetings, foreign language courses; $\hat{\delta}^{Language} = -0.48$), and spirituality (spiritual health, energy healing, nature, and spiritual support; $\hat{\delta}^{Spirituality} = -0.57$).

On the other hand, digitization enhances participation when activities are oriented towards business (business networking, innovation and start-ups, career growth; $\hat{\delta}^{Business} = 0.24$), science and technology (software development, coding, web design; $\hat{\delta}^{Tech} = 0.33$), financial investments and real estate ($\hat{\delta}^{Investments} = 0.11$), and gaming (board games, card games, and other gaming entertainment; $\hat{\delta}^{Games} = 0.33$).

These results suggest that the effects of digitizing community activities are more nuanced than past studies may have suggested. Business leaders are often presented with generic advice about organizing successful virtual events – such as considering whether the event involve a high level of one-on-one interaction, or whether the meeting goals are relationship-based or task-based (Clark 2021, Ringel 2021). The results of the Bayesian analysis complement these anecdotal suggestions with evidence that the suitability of digitization for community events may depend on factors that go above and beyond the levels of interactivity or the nature of the meeting goals in isolation. In addition to interaction requirements and task-based considerations, the interest-specific dynamics of an event are crucial to explain its attractiveness under increased digitization. For example, both socialization and business networking events require high levels of social interactions. Similarly, both a language class and a software workshop mostly pursue task-based goals. However, our results show that the effect of digitization on participants' engagement in one type of event is almost opposite to the other.

4.2 Counterfactual Analysis

After establishing that the effect of digitization varies across interest topics, we evaluate how much digitization affected the subset of events that were most probably digitized. We accomplish

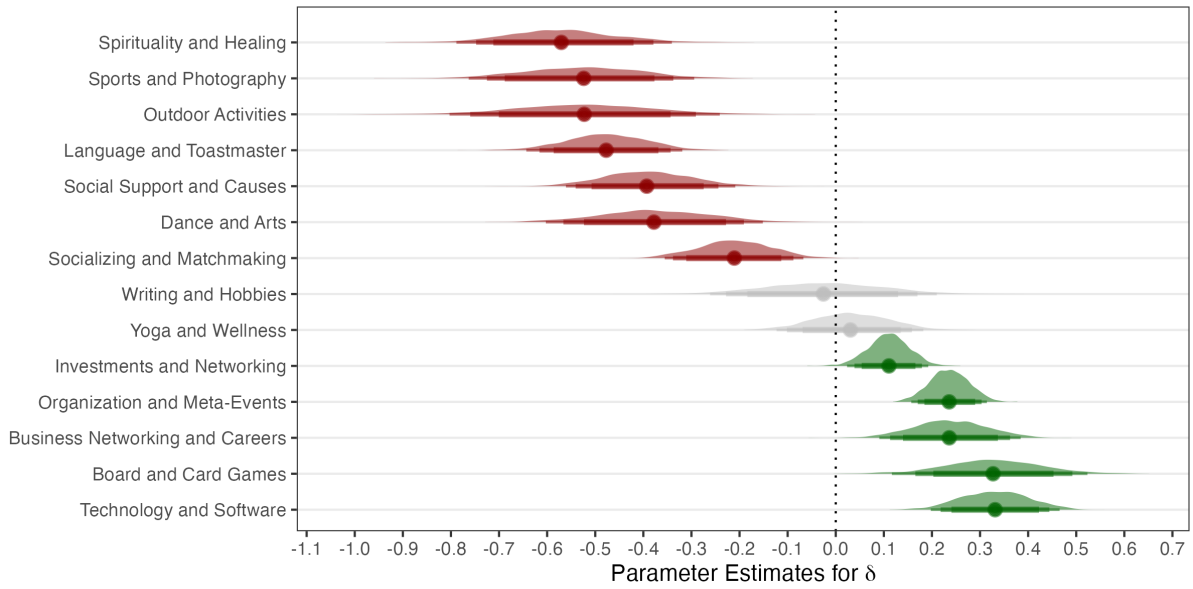


Figure 4: The Effect of Digitization is Highly Heterogeneous Across Interest Topics. *Note:* The figure shows the posterior density, mean, and 80%, 90%, and 95% Bayesian credible intervals of the δ estimates.

this through a counterfactual analysis, whereby we simulate responses to digitized events as if they were fully in-person, and again as if they were fully digitized. We then measure the difference in event participation under these fully digitized and in-person formats. In other words, we measure an average treatment effect on the treated (ATT).

We present the results of the policy evaluation in Figure 5. The counterfactual analysis reveals that digitization impacts not only *whether* attendees respond, but also *how* they respond to upcoming events in different interest categories. For instance, digitization significantly decreases the number of *positive* RSVPs to events related to music, dancing, arts, and culture (Arts and Music, average -10.15%. 95% distribution limits $[-25.2\%, -1.5\%]$), and writing clubs, skill development, religion, and hobbies (Writing and Hobbies, average -2.95% $[-12.8\%, -0.1\%]$). In other words, for an average-sized group organizing digitized events, digitization would cause a decrease of 63 potential attendees per event in the Arts and Music topic, and 28 attendees in the Writing and Hobbies topic. On the other hand, digitization increases the number of positive RSVPs to events related to business networking, innovation and start-ups, and career growth (Business Networking and Careers, average 9.42 $[-0.57\%, 30.6\%]$), and events related to financial investments and real estate (Investments and Networking, average 3.06 [0%, 10.17%]). On average, digitization would cause an increase of 51 potential attendees per event in the Business Networking and Careers topic, and an increase of 20 attendees in Investments and Networking.

The counterfactual analysis also provides additional insight into effect heterogeneity across events that share similar characteristics. For instance, digitization can have opposite effects

across goal-oriented meetings organized in different contexts: average positive participation in *writing* clubs decreases by 2.95% due to digitization, while average positive participation in *language* clubs increases by 2.72%.

To summarise, this counterfactual evaluation provides three important insights. First, it quantifies the heterogeneous effect of digitization on community participation in a situation where every other factor is held constant, except for the extent of event digitization. This allows us to understand and compare the effect of digitization across events that share similar characteristics, but differ in terms of context or purpose. Second, it suggests that inferences based solely on the ATT are, at a minimum, incomplete, and perhaps even misleading. Digitization has a heterogeneous, nuanced effect on community engagement with meaningful consequences for event management, with nearly opposite effects on prospective attendance depending on the type of event. Third, the analysis speaks directly to some of the most debated issues around digitization and in-person socialization. For example, business leaders, academics, and policy makers are currently discussing the implications and possible solutions to the "great mismatch" – a stark difference between employers and employees in their preference for in-person business activities (Robinson 2023). This analysis provides initial evidence that business networking meetings and career growth opportunities can continue to offer value to prospective participants, even when their formats are fully digitized, whereas other types of social activities may be better suited for in-person formats.

4.3 Robustness Checks

In this section, we assess the robustness of the main results to different model specifications.

4.3.1 Causal Random Forests

The parametric specification in Equation 3 assumes a functional form for expected utility and the unobserved error term ϵ_{ie}. In this section, we relax these assumptions and perform a non-parametric causal random forest (CRF) analysis (Athey and Wager 2019, Wager and Athey 2018). To train the CRFs, we used the observed RSVP choice, Y_{ie}, as the outcome variable; a binarized indicator for whether the probability of event digitization exceeded 50%, $\check{D}_e$, as the treatment indicator; and W_{ie}, W_g, and the city and week of the event as the deconfounding covariates. As mentioned earlier, we omit the group identifiers for computational reasons. However, following Athey and Wager 2019, the group identifiers are used to estimate the cluster-robust average treatment effects. Thus, the forests assume that the outcomes Y_{ie} of members in the same group may be arbitrarily correlated within a group. To improve precision, we followed

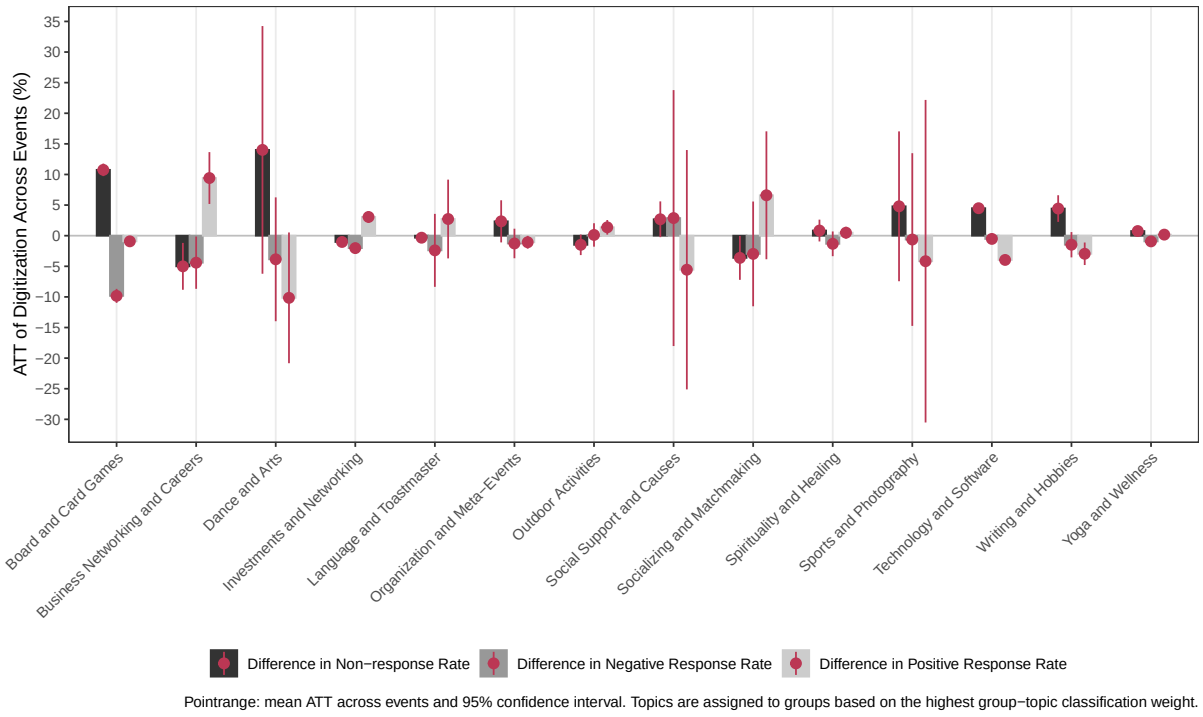


Figure 5: Average Percentage Changes in Counterfactual RSVP Values (ATT) Due to Digitization across Interest Topics. *Note:* Topics are assigned to groups based on highest group-topic classification weight.

Athey and Wager 2019 and trained two separate causal random forests. First, we trained a pilot random forest on all the covariates. Then, we trained a second forest using only the covariates with an above-average number of splits in the first forest. This approach allows the second forest to make more splits on the most important features in low-signal situations (Athey and Wager 2019).

Using an overlap-weighted average treatment effect estimator (Li et al. 2018), we find that the ATE of digitization in the training sample is -0.041 (95% CI $[-0.085; 0.003]$). However, as mentioned previously, this average effect potentially masks the considerable heterogeneity in this effect. Hence, we also calculate a CATT within each interest topic, per equation (12). The results, reported in Table 7, suggest that the counterfactual CATT estimates are largely robust to relaxing functional form assumptions: 13 of 14 counterfactual CATTs fall within the 95% CI of the random forest ATTs. The counterfactual effect that falls outside the random forest's confidence intervals refers to events organized in the Arts and Music category (with a counterfactual effect smaller than the effect estimated by the random forest). Estimates from the random forests show that the heterogeneous CATTs from the counterfactual analysis are largely robust to functional form assumptions for expected utility and the distribution of the unobserved error term.

	Causal Random Forests				Counterfactual SCM	
Topic	ATT	SD	2.5% CI	97.5% CI	Counterfactual ATT	Within CRF 95%CIs?
1 - Organization and Meta-Events	-0.006	0.101	-0.205	0.192	-0.018	Yes
2 - Social Support and Causes	0.293	0.430	-0.550	1.136	-0.119	Yes
3 - Language and Toastmaster	0.114	0.141	-0.162	0.390	0.002	Yes
4 - Yoga and Wellness	0.269	0.220	-0.161	0.699	-0.006	Yes
5 - Socializing and Matchmaking	-0.079	0.085	-0.246	0.088	0.020	Yes
6 - Investments and Networking	-0.085	0.112	-0.303	0.134	0.022	Yes
7 - Board and Card Games	-0.025	0.072	-0.167	0.117	-0.120	Yes
8 - Arts and Music	-0.035	0.134	-0.296	0.227	-0.354	No
9 - Spirituality and Healing	0.026	0.024	-0.021	0.074	-0.008	Yes
10 - Sports and Photography	-0.069	0.106	-0.278	0.139	-0.090	Yes
11 - Outdoor Activities	-0.096	0.126	-0.343	0.151	0.015	Yes
12 - Business Networking and Careers	-0.143	0.251	-0.636	0.350	0.042	Yes
13 - Writing and Hobbies	0.027	0.111	-0.190	0.244	-0.033	Yes
14 - Technology and Software	0.007	0.033	-0.058	0.071	-0.034	Yes

Table 7: Heterogeneous Treatment Effects on the Treated from the Counterfactual Analysis are Largely Robust to Functional Form Assumptions, and Fall Within the 95% Confidence Intervals of the ATTs from Causal Random Forests.

4.3.2 Nearest-Neighbor Matching

Next, we conduct a second robustness check that addresses potential concerns about covariate imbalance between digitized and in-person events. Specifically, we adjust members' exposure to digitized and non-digitized events non-parametrically, using nearest-neighbor matching of RSVPs on covariates (W_{ie} and W_g, Ho et al. 2011). We match covariates on the binarized treatment indicator $\check{D}_e$, described previously. The matching procedure results in a new sample, wherein members' assignment to digitized events is not associated with the covariates (Morgan and Winship 2015).

This approach has three advantages. First, it balances the identifying covariates across degrees of digitization. This is especially relevant for our empirical setting, where treatment is relatively uncommon. Second, matching allows for a comparison of RSVP decisions between digitized and non-digitized events, unconfounded by the measured and balanced covariates. Third, it helps us assess the robustness of the counterfactual CATTs. Appendix D provides additional details about the procedure and the matched sample.

Using the matched subset, we compute group-level CATTs for the same subset of groups considered in the counterfactual analysis. Then, we compare the distribution of group-level CATTs calculated on the matched subset with the distribution of counterfactual CATTs. Figure 6 presents the two distributions. Computing a distribution-free overlapping index between the distributions of counterfactual and matching CATTs, we find that the two distributions significantly overlap (overlapping index $\hat{\eta} = 0.60$, Pastore and Calcagnì 2019). Furthermore, the average group-level CATT estimates obtained from the counterfactual analysis and the matching procedure are not significantly different (Matching average = -0.004, SD = 0.6, 95% CI: [-0.14, 0.13]; Counterfactual average = -0.02, SD = 0.15, 95% CI: [-0.04, 0.01]. Paired

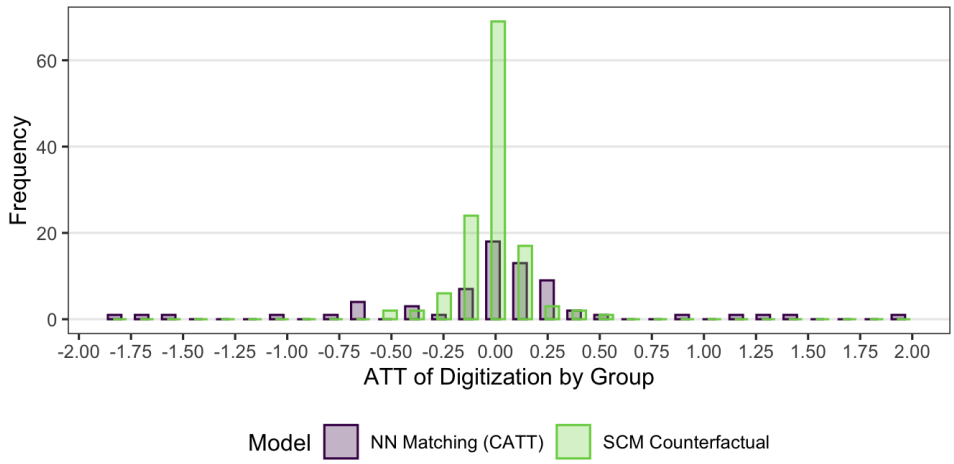


Figure 6: The Distributions of Group-level CATTs of Digitization Obtained from the Counterfactual Estimates and from the Matching Procedure Overlap. *Note:* N. treated groups = 126. The nearest-neighbor matching group-level CATT is calculated as the within-group average difference between outcomes under complete event digitization (Y_{ie}^{1}) and outcomes under complete in-person formats (Y_{ie}^{0}) on the matched data.

T-test: t = -0.04, df = 67).

In sum, both the CRF and matching analyses produce results that are broadly consistent with those obtained from the SCM.

5 Conclusions

Digitized social experiences are becoming more and more common. Although digitization can be a cheap and accessible alternative to in-person interactions, the lack of "human touch" may significantly impact the utility that people receive from these encounters. This is especially true in contexts of community-building, wherein the establishment of deep, trusting, social bonds is one of the main reasons for participation (Kang et al. 2014, Wang and Fesenmaier 2004). As business leaders and policy-makers evaluate the complex implications of digitizing local communities and their social encounters, the essential question is whether digitizing social events hinders people's engagement in their communities. More specifically: is digitization always a threat to social and community participation? Or does its impact vary across categories of interest and community topics?

In this study, we show that people are slightly less likely to participate in events, on average, if those events are digitized. However, the results also suggest that the impact of digitization on event participation intentions is highly heterogeneous. In particular, the results show that the central interest of the community is an important source of heterogeneity in how digitization affects community participation. For example, digitized events organized around entertainment

or socializing are less attractive to community members than their in-person counterparts. By contrast, digitized events for business networking, innovation and start-ups, career growth, science and technology, or gaming are at least as attractive – if not more attractive – than their in-person counterparts. These insights are robust to alternative model specifications.

From the perspective of future participants, the results suggest that *both* in-person and digitized activities can generate value, depending on the interest-specific characteristics of the events. These insights complement available anecdotal evidence on the suitability of digitization based on single event characteristics, such as the level of interactivity to be expected during the event, or the nature of the tasks that attendees will perform (Capossela 2022, Ringel 2021). Some of the empirical results are counter-intuitive with respect to these anecdotes. For instance, networking events should not be suitable for digitization as they require high levels of interactivity, but we find that they benefit from digitization when they are organized around career growth and business opportunities. This might suggest that there is no one-size-fits all solution when it comes to digitization, and that the presence or lack of "human touch" may not be an encompassing explanation for the success or failure of digitized activities. The results from this study, which capture diverse underlying interests and topic dynamics, can help event marketers and community organizers choose event formats based on the likelihood that they may benefit from digitization.

The opportunities and threats associated with digital transformation are also at the core of important policy discussions. Particularly, business leaders, academics, and policy-makers are concerned with the "great mismatch", resulting from diverging preferences for digitization between employers and employees (Robinson 2023). This study can provide initial support for companies looking to digitize part of their operations (such as company events, product demonstrations, and coding workshops), as well as initial quantitative insight into the potential benefits of digitizing business activities with social and networking components. Companies could also use the results from this study to segment or target consumers or employees, based on their preferences for different types of digitized events. Finally, this study constitutes an important quantitative benchmark for future research in the domain of digitization and its effects on social engagement.

This study presents several limitations, as well as avenues for future investigation. First, the study was conducted on data that precede the COVID–19 pandemic, which could limit the generalizability of the findings to a post-pandemic world. While acknowledging the limitation and the major disruptions caused by COVID–19, we argue that pre-COVID insights on the effects of event digitization can still be useful. Pre-COVID insights provide a baseline for future

comparison. By understanding how events were conducted and attended before the pandemic, managers and other researchers can benchmark any changes to event participation following COVID-19 disruptions and lockdowns. Additionally, pre-COVID data can help identify patterns in differential event participation that are likely to persist – or become even more relevant – in post-pandemic contexts. For example, fear and uncertainty regarding human-to-human virus transmission may have exogenously decreased the benefits from in-person interactions, exacerbating the positive effect of digitization in certain interest topics. On the other hand, the positive heterogenous effects of *not digitizing* social activities may be even stronger post-COVID, as the benefits from face-to-face meetings may be perceived as stronger and more urgent. Future studies should expand the evaluation of digitization policies in a post-COVID reality, and compare how the shift to remote working and fully digitized social activities affected the complex patterns of community participation.

A second limitation is related to the measurement of event participation and event digitization. In particular, the study relies on measures of participation intentions in place of actual attendance. RSVPs are often used by event organizers as a proxy for measuring intention to participate and are correlated with actual attendance. Thus, they can still provide valuable insights into people's intentions and interests in the groups and their events. However, RSVPs may not always accurately predict actual attendance, which could have implications for the magnitude of the observed effects. On the one hand, a *positive* treatment effect of digitization on RSVPs may be *overestimating* the effect that digitization would have on actual attendance. RSVPing to an event requires less effort than attending it, and no-shows may be relatively less problematic for a digitized event than for an in-person event. Meetup group organizers already acknowledge the problem of no-shows and try to limit their occurrence, for example by imposing penalties on repeated no-shows, or by requiring a financial commitment at the time of RSVP creation. On the other hand, *negative* treatment effects of digitization on RSVPs may be *underestimating* the effect of digitization on attendance. If digitization is unattractive for prospective attendees, attendance rates could be even lower than RSVP rates. However, in the case of negative effects, both RSVPs and attendance rates cannot be lower than zero, which may limit the extent of underestimation. To overcome these limitations, future research may study the effect of event digitization on additional outcomes – such as actual attendance, passive participation, and even negative or disruptive participation (Ardichvili et al. 2003, Dutta-Bergman 2005, Brodie et al. 2013, Kang et al. 2014). More specifically, it would be interesting to assess if members exploit the increased anonymity from digitization to be more disruptive, or to choose negative forms of engagement.

Third, this study measures event digitization using point estimates based on unstructured text data, rather than a "ground truth" variable describing the exact format of the event. This limitation is especially important when the continuous digitization variable is transformed into a binary treatment indicator, using the 50% probability estimated by the Support Vector Machines as a dichotomization threshold. In particular, the binary treatment indicator may be overestimating (or underestimating) the incidence of digitization for events just above (or below) the 50% threshold. At the time of data collection, a "ground truth" variable for event format was unavailable. However, Meetup recently changed its event creation interface, and is allowing organizers to indicate that an event is happening online, and to include a featured link to the online meeting (Weger and Meetup.com 2021). Future studies could take advantage of this change in the platform, and rely on the new self-reported measure of digitization, in combination with unstructured text, to estimate treatment effects of digitization on event participation.

Finally, while the study identified several event categories in which digitization had a significant impact on participation, it did not explore the specific mechanisms through which digitization influenced participation. The results could be consistent with the influence of several drivers, such as the degree of "human touch" and social presence that members expect to find in community events, differences in access to digital resources and digital literacy, or the extent to which fully digitized events are perceived as attractive on a platform where in-person events are the norm. Future research could address this limitation by (i) conducting studies in diverse contexts, including different platforms; (ii) examining other member-level moderators, including satisfaction and familiarity with digitization technologies, or expectations about "human touch" in community interactions; and (iii) running field experiments, with random assignment of attendees to digitized versus in-person event formats, to disentangle potential mechanisms. We encourage future research in field settings to carefully consider the ethical concerns that arise from limiting prospective attendees to access only a given event format. These concerns are especially salient in the aftermath of a global pandemic, as preventing attendees from accessing a virtual event might cross an ethical boundary. Such ethical considerations are also salient for event categories that generate substantial socioeconomic value for prospective attendees, such as career networking and social support groups, as the exclusion from an in-person meeting could have serious consequences for attendees' personal and financial well-being. Conditional on resolving these ethical issues, field experiments could provide valuable information about the boundaries and dynamics of the digitization effects.

The findings and suggestions from this study represent initial quantitative and generalizable insights into how digitization technologies can enhance social engagement in local and

distributed communities. To our knowledge, this study is the first to address several open issues in the study of digitization with extensive data and diverse methodologies. In particular, we integrate structured and unstructured data with natural language processing to provide a benchmark measure of the digitization of social experiences. Additionally, we use both structural causal models and machine learning methods to offer a robust and comprehensive understanding of the implications of increased digitization. We hope to inspire future research to further investigate, question, and ultimately expand the boundaries of human connection in an increasingly digitized world.

6 Declarations

6.1 Funding and Competing Interests

All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript. The authors have no funding to report.

References

6Connect (2020) Virtual Events vs. Physical Events: The Definitive Pro-Con List. URL https://www.6connex.com/blog/virtual-events-vs-physical-events-pro-con-list/.

Aarons-Mele M (2022) Working Through Your On-Camera Meeting Anxiety URL https://hbr.org/2022/07/working-through-your-on-camera-meeting-anxiety.

Algesheimer R, Borle S, Dholakia UM, Singh SS (2010) The impact of customer community participation on customer behaviors: An empirical investigation. *Marketing science* 29(4):756–769.

Ardichvili A, Vaugh P, Wentling T (2003) Motivation and barriers to participation in virtual knowledge-sharing communities of practice. *Journal of Knowledge Management* 7(1):64–77.

Arun R, Suresh V, Veni Madhavan C, Narasimha Murthy M (2010) On finding the natural number of topics with latent dirichlet allocation: Some observations. *Advances in Knowledge Discovery and Data Mining: 14th Pacific-Asia Conference, PAKDD 2010, Hyderabad, India, June 21-24, 2010. Proceedings. Part I 14*, 391–402 (Springer).

Atasoy O, Morewedge CK (2018) Digital goods are valued less than physical goods. *Journal of consumer research* 44(6):1343–1357.

Athey S, Wager S (2019) Estimating treatment effects with causal forests: An application. *arXiv preprint arXiv:1902.07409* .

Bettinger EP, Fox L, Loeb S, Taylor ES (2017) Virtual classrooms: How online college courses affect student success. *American Economic Review* 107(9):2855–2875.

Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. *Journal of machine Learning research* 3(Jan):993–1022.

Bonsall A (2022) 3 Types of Meetings — and How to Do Each One Well URL https://hbr.org/2022/09/3-types-of-meetings-and-how-to-do-each-one-well.

Bourreau M, Doğan P (2018) Gains from digitization: Evidence from gift-giving in music. *Journal of Economic Behavior & Organization* 149:106–122.

Brodie RJ, Ilic A, Juric B, Hollebeek L (2013) Consumer engagement in a virtual brand community: An exploratory analysis. *Journal of business research* 66(1):105–114.

Cao J, Xia T, Li J, Zhang Y, Tang S (2009) A density-based method for adaptive lda model selection. *Neurocomputing* 72(7-9):1775–1781.

Capossela C (2022) To Get People Back in the Office, Make It Social URL https://hbr.org/2022/09/to-get-people-back-in-the-office-make-it-social.

Carpenter B (2023) 39 Event Invite Subject Lines: Stand Out amp; Increase Attendance. *Eventbrite Blog* URL https://www.eventbrite.com/blog/event-invite-subject-lines-ds00/#Subject-lines-that-encourage-RSVPs.

Catapano R, Shennib F, Levav J (2022) Preference reversals between digital and physical goods. *Journal of Marketing Research* 59(2):353–373.

Clark D (2021) How to Host a Virtual Networking Event. URL https://hbr.org/2020/05/how-to-host-a-virtual-networking-event.

Cohn A, Gesche T, Maréchal MA (2022) Honesty in the digital age. *Management Science* 68(2):827–845.

Dessart L, Veloutsou C, Morgan-Thomas A (2015) Consumer engagement in online brand communities: a social media perspective. *Journal of Product and Brand Management* 24(1).

Deveaud R, SanJuan E, Bellot P (2014) Accurate and effective latent concept modeling for ad hoc information retrieval. *Document numérique* 17(1):61–84.

Dutta-Bergman MJ (2005) The antecedents of community-oriented internet use: Community participation and community satisfaction. *Journal of Computer-Mediated Communication* 11(1):97–113.

Eventbrite (2023) 5 minutes to better attendance: How to promote events on Facebook. *Eventbrite Blog* URL https://www.eventbrite.com/blog/power-of-facebook-events-ds00/.

Gratton L (2021) How to Do Hybrid Right URL https://hbr.org/2021/05/how-to-do-hybrid-right.

Griffiths TL, Steyvers M (2004) Finding scientific topics. *Proceedings of the National academy of Sciences* 101(suppl_1):5228–5235.

Ho DE, Imai K, King G, Stuart EA (2011) MatchIt: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software* 42(8):1–28, URL http://dx.doi.org/10.18637/jss.v042.i08.

Kang J, Tang L, Fiore AM (2014) Enhancing consumer–brand relationships on restaurant facebook fan pages: Maximizing consumer benefits and increasing active participation. *International Journal of Hospitality Management,* 36:145–155.

Lee S, Lee JH, Jeong Y (2022) The effects of digital textbooks on students' academic performance, academic interest, and learning skills. *Journal of Marketing Research* 00222437221130712.

Li F, Morgan KL, Zaslavsky AM (2018) Balancing covariates via propensity score weighting. *Journal of the American Statistical Association* 113(521):390–400.

Luangrath AW, Peck J, Hedgcock W, Xu Y (2022) Observing product touch: The vicarious haptic effect in digital marketing and virtual reality. *Journal of Marketing Research* 59(2):306–326.

Meetupcom (2018) How do I manage attendance for an event? URL https://help.meetup.com/hc/en-us/articles/360002870552-How-do-I-manage-attendance-for-an-event-.

Morgan SL, Winship C (2015) *Counterfactuals and causal inference* (Cambridge University Press).

Murzintcev N (2020) *ldatuning: Tuning of the Latent Dirichlet Allocation Models Parameters*. URL https://CRAN.R-project.org/package=ldatuning, r package version 1.0.2.

Pastore M, Calcagnì A (2019) Measuring distribution similarities between samples: a distribution-free overlapping index. *Frontiers in psychology* 10:1089.

Riedl C, Kim YJ, Gupta P, Malone TW, Woolley AW (2021) Quantifying collective intelligence in human groups. *Proceedings of the National Academy of Sciences* 118(21):e2005737118.

Ringel R (2021) When Do We Actually Need to Meet in Person? URL https://hbr.org/2021/07/when-do-we-actually-need-to-meet-in-person.

Robinson B PhD (2023) 'The Great Mismatch': Employers Firmer On Return-To-Office Policies In 2023 URL https://www.forbes.com/sites/bryanrobinson/2023/01/01/the-great-mismatch-employers-firmer-on-return-to-office-policies-in-2023/?sh=d8d2ddbbe1f3.

Rosenbaum PR, Rubin DB (1983) The central role of the propensity score in observational studies for causal effects. *Biometrika* 70(1):41–55.

Sconti A (2022) Digital vs. in-person financial education: What works best for generation z? *Journal of Economic Behavior & Organization* 194:300–318.

Spinks D (2020) Tools you can use to host virtual conferences and events. URL https://web.archive.org/web/20200422202751/https://twitter.com/DavidSpinks/status/1235993594705534976.

Stan Development Team (2021) Stan Modeling Language Users Guide and Reference Manual, 2.27. URL https://mc-stan.org/.

Talts S, Betancourt M, Simpson D, Vehtari A, Gelman A (2018) Validating bayesian inference algorithms with simulation-based calibration. *arXiv preprint arXiv:1804.06788* .

Tang A (2023) The perfect event RSVP email Templates. *Eventbrite Blog* URL https://www.eventbrite.com/blog/the-perfect-event-rsvp-email-templates/.

The CMO Survey (2021) CMO Survey. URL https://cmosurvey.org/results/26th-edition-february-2021/.

Tibshirani J, Athey S, Wager S (2021) *grf: Generalized Random Forests*. URL https://github.com/grf-labs/grf, r package version 1.2.0.0.

Touré-Tillery M, Wang L (2022) The good-on-paper effect: How the decision context influences virtuous behavior. *Marketing Science* .

US Census Bureau (2010) American community survey. URL https://is.gd/3wGOwg.

Wager S, Athey S (2018) Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association* 113(523):1228–1242.

Wang Y, Fesenmaier DR (2004) Towards understanding members' general participation in and active contribution to an online travel community. *Tourism management* 25(6):709–722.

Weger B, Meetupcom (2021) Product Updates for Online Events (Part 1). URL https://www.meetup.com/blog/product-updates-for-online-events-part-1/.

Wiegand N, Imschloss M (2021) Do you like what you (can't) see? the differential effects of hardware and software upgrades on high-tech product evaluations. *Journal of Interactive Marketing* 56(1):18–40.

Wiertz C, de Ruyter K (2007) Beyond the call of duty: Why customers contribute to firm-hosted commercial online communities. *Organization studies* 28(3):347–376.

Yokoi T, Obwegeser N, Beretta M (2021) How Digital Inclusion Can Help Solve Grand Challenges URL https://sloanreview.mit.edu/article/how-digital-inclusion-can-help-solve-grand-challenges/.

A Event Awareness Indicator

In constructing the estimation sample, we make a distinction between members who were potentially unaware of an event, and members who were potentially aware. We distinguish the two types of members with an indicator variable.

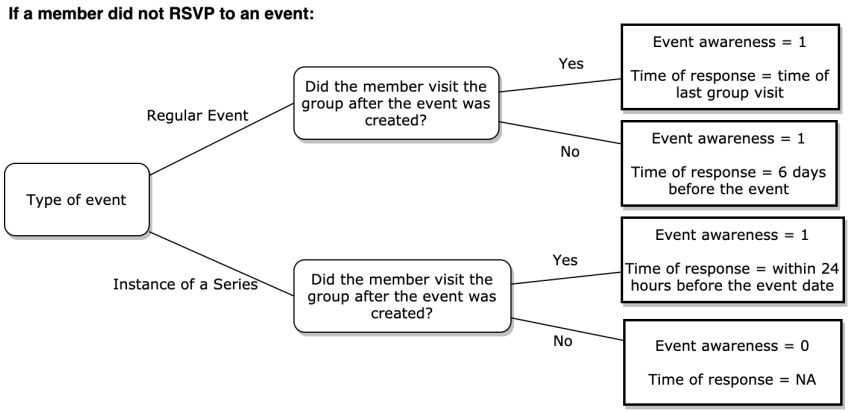
To construct the awareness indicator, we relied on two variables available for each member-event pair in our sample: the *event series* indicator and the *timestamp of last group visit*. The event series indicator is set to 1 if the event is part of an event series – a set of events that repeat with fixed frequency (every week, every two weeks, or every month). The *event series* indicator is set to 0 if the event is a standard (non-repeating) event.

For standard events, we also exploited a feature active on Meetup in 2019. In 2019, Meetup sent RSVP reminders via email to all group members 6 days before the scheduled event date. Because of the reminder, we set the event awareness variable to 1 for all members who did not RSVP to a standard event. We also imputed their potential time of response, and set it at a date corresponding to 6 days before the scheduled event date. If a member who did not RSVP to a regular event visited the group after the event creation date and before the 6-day threshold, then we imputed the time of response as the most recent time at which the member visited the group.

For event series, if a member did not RSVP to an event, but visited the group after the event was created, we assumed that this member *decided* not to RSVP to the event. Therefore, we assumed that the person was aware of the existence of the event, and set the event awareness variable to 1. We also imputed their potential time of response, and set it at a date corresponding to 24 hours before the event occurrence. This is the last time window in which the member could have made a decision about RSVPing.

Finally, if a member did not RSVP to a recurring event, and did not visit the group before the event creation date, then we assumed that this member was potentially unaware of the event. We set the event awareness variable to 0, and did not impute their potential time of response. Figure 7 summarises the event awareness measurement procedure.

Figure 7: Event Awareness Indicator – Decision Flow



B Measuring Event Digitization

The raw data from the Meetup API do not include a field describing event formats (i.e. online, in-person, hybrid). We therefore define a measure of event digitization based on other information in the raw data. As a source of digitization information, we rely on the free-text *event descriptions* created by the event organizers. Event descriptions are visible to group members, and are crafted to inform perspective attendees about the event format (online, offline, or hybrid). Additionally, descriptions typically provide details on how to find the event location, or which activities will be performed for the duration of the event. In practice, group members can use the event descriptions to evaluate the event attractiveness, and form a decision about their event participation intentions. We used the *event venue* field as an additional source of information. The event venue field typically contains the address of the location in which in-person events take place, or the name of the platform used to host digitized events.

To extract information from the text, we created a list of non-empty event descriptions, using the events organized by groups in the sample between the group's inception and June 2019. We processed the text to remove HTML tags, trailing whitespace, English stopwords, phone numbers, punctuation, and special characters. We then used a subset of the cleaned description text as an input for two support vector machine (SVM) classifiers.

B.1 Construction of the Training Subset

To train the SVM classifiers, we created a training set of descriptions labeled as "online", "offline", or "both" in three steps.

B.1.1 Step 1: Keyword-matching Training Cases

After cleaning the event description texts, we performed a keyword-matching task to label online and offline events based on the words they mentioned.

We defined three vectors of keywords that could potentially indicate that an event was completely digitized or completely offline. The first vector included six digitized-event keywords: "online event", "remote meeting", "webinar", "gotomeeting", "webcast", and "remotely". We also obtained a list of common tools used to organize digitized community activities from Spinks 2020, and appended the list of tools to the digitized-event keyword vector.

The second vector included ten keywords related to event locations that strongly suggest a completely digitized event format: "http://", "https://", "online", "computer", "webinar", "anywhere", "your house", "iphone", "webcast", and "your computer".

The third vector included twelve offline-event keywords: "space provided", "breakfast served", "coffee served", "seats", "snacks", "drink", "drinks", "meet greet", "doors", "indoor", "outdoor", and "entrance".

We labeled the event descriptions that contained the offline-event keywords as "offline" events. We excluded any event descriptions that contained any of the digitized-event or digitized-location keywords from the "offline" cases.

Then, we labeled as "online" the cases that contained both the digitized-event keywords and the digitized-location keywords, and excluded any descriptions that contained the offline-event keywords.

Finally, we excluded from the training set all the cases that were labeled as offline if the event venues contained the keywords "http://", "https://", "online", "computer", "webinar", "anywhere", "your house", "iphone", "webcast", or "your computer".

We manually checked that the cases labeled as "online" were accurately classified based on the text descriptions. With this procedure, we labeled 157 cases as "online".

B.1.2 Step 2: Random Sampling and Human Ratings

In addition to the cases labeled using keyword-matching, we drew a random sample of 3000 event text descriptions that were manually labeled by two research assistants (RAs). The RAs classified the events depending on whether the text descriptions were describing activities with a "Digital/Virtual" format (i.e., people in the group met in a digitized, digital, online activity), and/or an 'In-Person" format (i.e., people in the group met face-to-face during the activity).

The RAs labeled the event descriptions with the class that most appropriately described the activity format ("Digital/Virtual" and/or "In-Person"). The events could be labeled as both "Digital/Virtual" and "In-Person" – in that case, the activity would be typically described as "Hybrid". When the two RAs chose different classifications for the same description, a third rater (who was not previously involved in the classification task) resolved the disagreements. The labeling phase resulted in a data set of 2850 labelled cases, of which 158 were classified as "Digital/Virtual", 2679 as "In-Person", and 14 as both ("Hybrid").

150 of the 3000 text descriptions were either empty, unintelligible, or written in a non-English language, and therefore were not classified by the RAs. We appended the 157 cases labeled with keyword-matching to the training cases labeled by the RAs.

B.1.3 Step 3: LDA Topic Model

The use of keyword-matching could potentially introduce bias in the measurement of event digitization, for example, due to the popularity of certain tools at the time of data collection. To address and mitigate the selection concerns, we trained an LDA topic model with 32 topics on all the available event descriptions.

The word-per-topic probabilities suggested that one of the topics (topic 9, including "zoom", "online" and "https" in the top 30 word-per-topic) included events with digitized formats. Therefore, we added all events with high probability of belonging to topic 9 to the training set.

After removing duplicates, the training set included 3033 labeled cases: 172 labelled as "Digital/Virtual", 2693 as "In-Person", and 14 as both "Digital/Virtual" and "In-Person".

Table 8: SVM 10-fold Cross-Validated Prediction Accuracies

	Prediction Accuracy (%)	
CV Fold	In-Person SVM	Digital/Virtual SVM
1	97.67	97.28
2	99.27	98.90
3	98.15	97.78
4	97.86	97.86
5	97.98	96.64
6	98.93	97.52
7	98.01	98.34
8	99.29	98.23
9	98.98	98.31
10	99.34	99.01
Average	98.55	97.99

B.2 SVM Predictions

We trained two Support Vector Machines (SVMs) on the set of labeled cases. We trained the first SVM using the "In-Person" label, and the second SVM using the "Digital/Virtual" label. Then, we let the two SVMs predict the most likely class of all the remaining unlabeled event descriptions (respectively "In-Person" versus "Not In-Person", and "Digital/Virtual" versus "Not Digital/Virtual"). This prediction step resulted in four new variables for each event description: (1) "In-Person" prediction label: most likely class of the text description ("In-Person" or not "In-Person") based on the SVM model trained on the "In-Person" label; (2) Probability associated with the "In-Person" (or not "In-Person") predicted class; (3) "Digital/Virtual" prediction label: most likely class of the text description ("Digital/Virtual" or not "Digital/Virtual") based on the SVM model trained on the "Digital/Virtual" label; (4) Probability associated with the "Digital/Virtual" predicted class.

B.3 SVM Performance

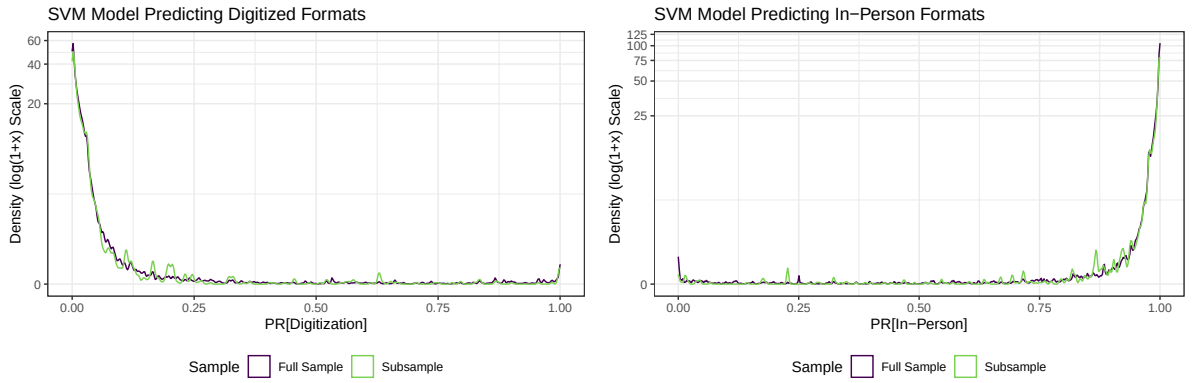
Using 10-fold cross-validation, the two SVM models achieved between 96.6% and 99.3% prediction accuracy. Table 8 reports the 10 prediction accuracies resulting from the cross-validation for each of the two models. The "In-Person" model achieved an average 98.5% prediction accuracy, while the "Digital/Virtual" model achieved an average 97.9%.

B.4 Prediction Descriptive Statistics

Figure 8 shows the distribution of event format probabilities predicted by the two SVM models (recall one model predicts digitization, the other in-person events). Panel (a) shows the distribution of probabilities that events are digitized (versus not digitized); Panel (b) shows the distribution of probabilities that events are in-person (versus not in-person). Both panels reflect a setting where in-person events are the norm. Furthermore, Figure 8 suggests the distributions of predictions in the estimation sample closely match the distribution in the full data set.

Table 9 describes which labels were attributed to each event in the panel. The vast majority of the events (99.6% of the total) were labeled consistently across prediction models. A small fraction of events

Figure 8: Distributions of Prediction Accuracies from Digital/Virtual SVM Model (a) and In-Person SVM Model (b)



(a) (b)

Table 9: Event Classification Labels from the In-Person and the Digital/Virtual SVM Models

Predicted Label				
In-Person SVM	Digital SVM	N	Total Events	(%)
Not In-Person	Not Digital/Virtual	900	562061	0.002
In-Person	Digital/Virtual	1397	562061	0.002
Not In-Person	Digital/Virtual	4236	562061	0.008
In-Person	Not Digital/Virtual	555528	562061	0.988

(0.4%) were labeled differently by each SVM model – the 0.2% of the event was labeled as both "In-Person" and "Digital/Virtual", and the 0.2% was labeled as neither. Inspecting a random sample of event descriptions, the inconsistent labels can be explained in three ways. One type of inconsistency derives from a misclassification – one of the two labels is correct, and the other is incorrectly classified. In this case the prediction accuracies are informative, and the label with highest prediction accuracy is typically the right one. The second type of inconsistency derives from events that actually have blended formats. These events are typically in-person, but offer a virtual live stream, real-time videos, or asynchronous digital material. The last type of inconsistency describes events with little or no information, and reflects the low classification confidence of either or both SVM models.

Finally, figure 9 shows that, overall, the predictions from the two models appear highly correlated, and that the majority of the predictions are concentrated in the upper-left (Not In-Person, Digital/Virtual) and bottom-right (In-Person, Not Digital/Virtual) regions of the plot.

Table 10 reports the number of events classified by the two SVM models in the full dataset (columns 1-3), and the number of members exposed to different combinations of event formats (columns 4-6).

Figure 9: Prediction Accuracies from In-Person SVM Model (x) and Digital/Virtual SVM Model (y). The correlation coefficient between x and y is equal to -0.908 ($t = -1627.5, df = 562059$, p-value < 0.001)

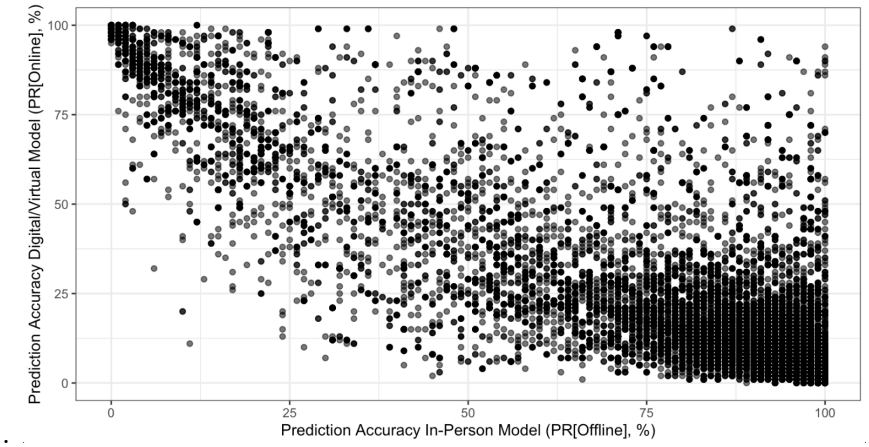


Table 10: Number of Events Classified by Format (columns 1-3); and Number of Members Exposed to Different Combinations of Event Formats (columns 4-6).

	N. Events			N. Members Exposed to Format		
	Digitized/Not In-Person	In-Person/Not Digitized	Both Formats	Digitized/Not In-Person	In-Person/Not Digitized	Both Formats
Model	(1)	(2)	(3)	(4)	(5)	(6)
SVM Digitization	1454	116586	286	555	261811	23364
SVM In-Person	1578	116512	236	733	260135	24862

Note: the classification into "digitized", "in-person", and "both" event class was performed using a 50% threshold for the predicted SVM probabilities ("digitized" if PR[Online] $>$ 50%; "in-person" if if PR[Online] $<$ 50%, and "both" if PR[Online] = 50%). The numbers refer to the full datset.

C LDA Topic Modeling of Interest Categories

We performed LDA topic modeling to classify events into latent interest topics, based on the free text descriptions provided by the event organizers (Blei et al. 2003). This analysis has the objective of reducing the number of interest categories describing the events, starting from an upper bound of 34 original interest categories included in the raw Meetup data. Reducing the number of interest categories will then help mitigating the problem of sparsity in the distribution of digitization probabilities across interest categories, as many of the 34 interest categories have low proportions of digitized events (see Section 2.7 for more details).

We select number of latent topics for LDA models based on 4 metrics from Arun et al. 2010, Cao et al. 2009, Deveaud et al. 2014, Griffiths and Steyvers 2004, using the R package "lda tuning" (Murzintcev 2020). We chose a 14-topic model solution, based both on the quantitative criteria and on achieving the objective of dimension reduction (Figure 10). In particular, we noted that the Griffiths and Steyvers 2004 metric has an inflection at 14 topics, and then plateaus. It is also the only value of the Griffiths and Steyvers 2004 metric to overlap with the metric from Deveaud et al. 2014, which in turn drops sharply at 15 topics. We note similar patterns for the Cao et al. 2009 metric: the value of the metric drops sharply at 14 topics, and then increases. The nearest drops are only recorded at 18 and 23 topics. However, a larger number of topics would not help with mitigating the sparsity problem.

Figure 11 shows the top 30 words associated with each of the 14 topics. These per-topic-per-word probabilities help us to qualitatively characterise the topics in Section 2.7, and to interpret the main results of this study.

Figure 10: Recommended Topic Solutions

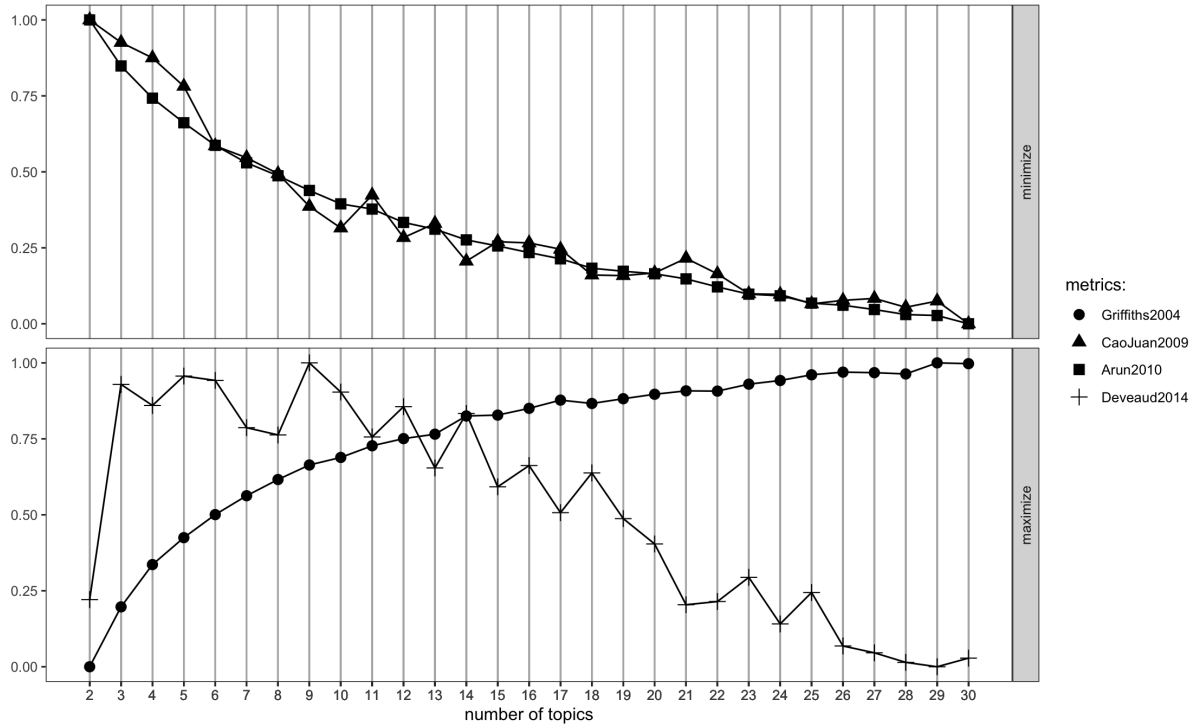
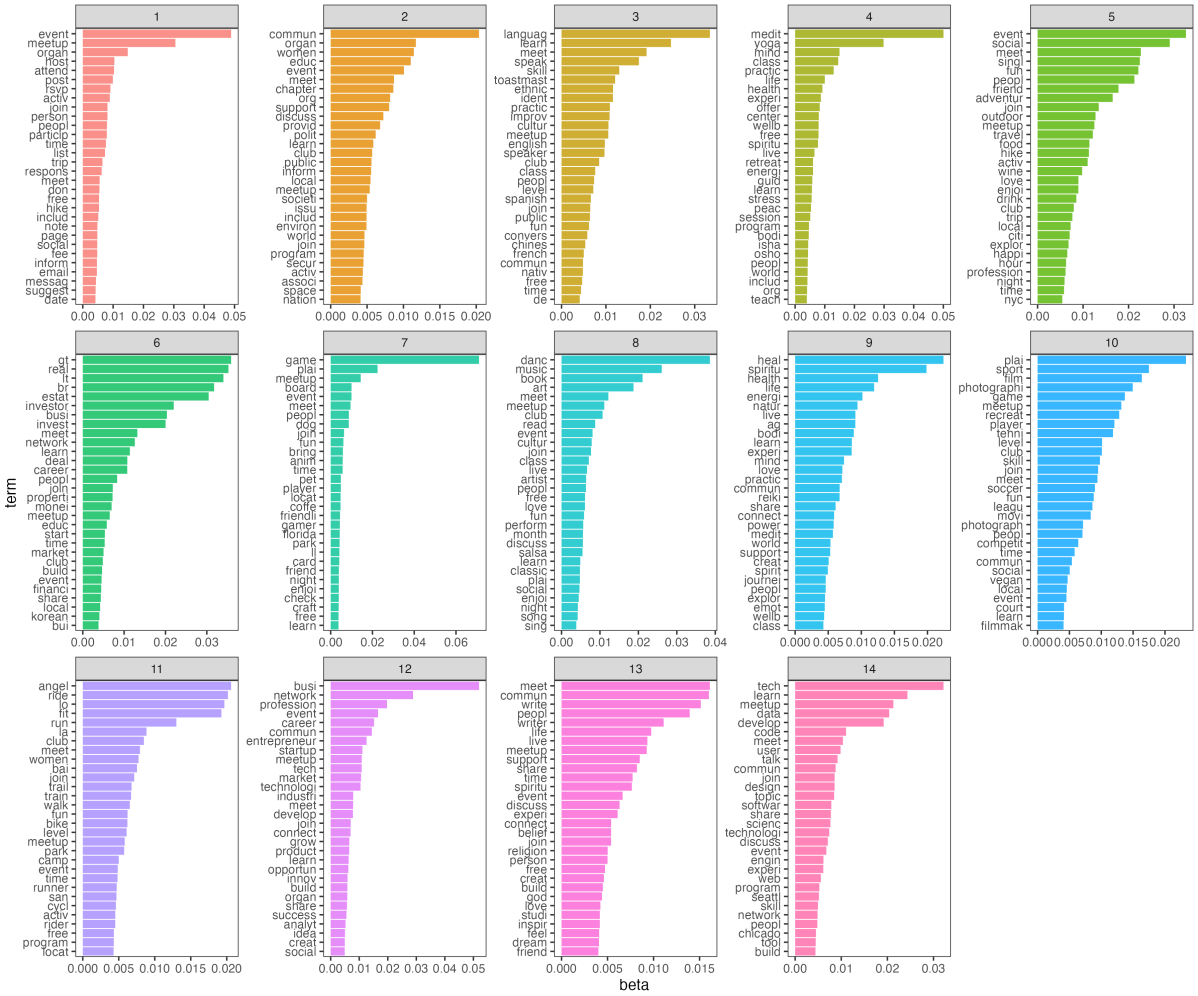


Figure 11: 14-Topic LDA Solution per-topic-per-word Probabilities (β)



D Nearest-Neighbor Matching on Covariates

We used the MatchIt package in R (Ho et al. 2011) to pair, subset and sub-classify observations, and create groups of RSVPs exposed to digitized versus non-digitized events that are balanced on identifying pre-treatment covariates $(W_{ie}, W_g, \tau_t, \zeta_m, \eta_g)$. For the matching procedure, we use a binary treatment indicator taking value 1 when the probability of event digitization exceeds 50%. The output of the MatchIt algorithm is a matched subset of observations, balanced on $W_{ie}, W_g, \tau_t, \zeta_m, \eta_g$.

The matched set includes 7,587 control observations, and 65,352 treated observations. Figure 12 shows the distributions of the average of the absolute differences of propensity scores between observation pairs ("distance"). The distance is larger if two units have very different estimated propensity scores. Therefore, in case of covariate imbalance between the treated and control samples, the two distributions do not overlap. The left panel in Figure 12 shows that, before matching, the distributions of the distance between treated and control units did not overlap, which suggests covariate imbalance. The right panel demonstrates that the matching procedure improved the distributional balance.

Figure 12: Density Plot for the Distributional Balance between Treatment and Covariates.