Social Science Research xxx (xxxx) xxx

Contents lists available at ScienceDirect

Social Science Research

journal homepage: http://www.elsevier.com/locate/ssresearch

Emergence of diverse and specialized knowledge in a metropolitan tech cluster

Daniel DellaPosta [a,*], Victor Nee [b]

[a] Department of Sociology and Criminology, The Pennsylvania State University, United States
[b] Department of Sociology, Cornell University, United States

ARTICLE INFO

Keywords:
Social networks
Tech economy
Specialization
Knowledge
Social media

ABSTRACT

Specialized knowledge is increasingly central in modern information- and technology-oriented economies, yet we know surprisingly little about how this knowledge is organized. We trace the evolution of specialized knowledge at both the individual- and network-levels by analyzing email exchanges shared among members of a large tech professional community in New York City over seven years. We find a shift over time toward the emergence of an increasingly specialized ecology of knowledge and information. This division of knowledge is driven by the influx of new cohorts of participants with different knowledge and interests than those already there. Yet, even as individual contributors increasingly sort into specialized niches, the community as a whole remains robust in its ability to address topics of diverse concern. This study illustrates how new sources of data enable us to see with greater clarity the structures underpinning modern knowledge-based innovation clusters.

1. Specialization and the division of knowledge

From computing to biology, scholars have long noted a tendency for systems across a wide variety of domains to evolve from lesser to greater states of internal complexity and differentiation (Simon, 1981). This includes the networks of relationships linking together human actors engaged in economic activity, as theorized by both classical (Durkheim, 1933; Smith, 1776) and contemporary (Jackson, 2008; Mani and Moody, 2014; Padgett and Powell, 2012) observers in the social sciences. In industrial clusters, specialization increasingly refers not just to physical capabilities or points in an assembly line, but rather to agglomerative growth of available information and knowledge (Bell, 1973; Castells, 1996; Mokyr, 2002; Powell and Snellman 2004; Saxenian, 1994; Varian, 1995). This "division of knowledge" underlying the division of labor in modern information- and technology-oriented economic spheres allows for diverse bits of information and expertise to combine, producing new and sometimes unexpected innovations (Glaeser, 1999).

Yet the shift toward increasing reliance on specialized knowledge implies a paradox: individual actors must increasingly hold *specialized* expertise in one domain while the collective group must also be able to access an increasingly *diverse* pool of knowledge workers, and overall knowledge and information. Are these tendencies toward both specialization and diversification conflictual or complementary? Does individual specialization lead to balkanization in which communities of specialists become sealed off from one another, or does it provide a foundation for integrating diverse knowledge and expertise?

Surprisingly little attempt has been made to answer these questions at either the micro-level of individual human actors

* Corresponding author. 503 Oswald Tower, The Pennsylvania State University, University Park, PA, 16802, United States.
E-mail addresses: djd78@psu.edu (D. DellaPosta), victor.nee@cornell.edu (V. Nee).

https://doi.org/10.1016/j.ssresearch.2019.102377
Received 25 March 2019; Received in revised form 24 September 2019; Accepted 16 October 2019
Available online 18 October 2019
0049-089X/

participating in these industrial spheres or the macro-level of industrial communities. One line of previous research on knowledge economies has focused on analyzing the ways in which formal contracts between firms provide network conduits for information gathering, knowledge spillover, and innovation (e.g. Owen-Smith and Powell, 2004; Whittington et al., 2009). This work usefully highlights the increasing centrality of knowledge and expertise held by individuals and circulated through network ties. In this framework, the hard boundaries between firms are less consequential than the networks linking firms together in an ecology of knowledge entrepreneurs. Rather than a producer of knowledge, the firm provides an institutional setting for the integration of diverse knowledge held by the individuals contributing to the firm's activities (Grant, 1996). Yet, previous work—which focuses on organizational-level processes—has not directly analyzed or measured the evolution of industrial clusters as the evolution of networks of *individuals* with varying bases of knowledge, expertise, and knowhow. To do so requires a different conceptual and methodological framework than is possible within the boundaries of firm-level studies.

From a different vantage point, economists have used education and other measures of human capital to study the pool of specialized knowledge available for use in particular industries and regions. However, as Hidalgo points out, "schooling is certainly not a great proxy for knowhow and knowledge, since it is by definition a measure of the time spent in an establishment, not of the knowledge embodied in a person's brain" (Hidalgo, 2015, p. 149). Similarly, individual-level measurements tell us little about the mix of specialized knowledge that workers in a regional technological economy such as Silicon Valley possess). A population of knowledgeable individuals can feature a meager overall pool of knowledge and expertise if all of those individuals know only the same things. In contrast, a population of narrower specialists can nevertheless present an impressive array of knowledge when the bits held by each individual differ, while complementing those held by others.

To explore the structure and evolution of knowledge specialization, we employ new strategies that allow us to track information-sharing activity among tech-startup professionals during the emergence and rapid growth of New York City's tech economy (e.g. Cometto and Piol, 2013). The analysis is based on publicly archived records of participation in a hybrid online-offline group called the New York Tech Meetup, currently boasting over 60,000 members, which is at the social and organizational epicenter of this booming tech-startup economy. Hosted on Meetup.com, the Tech Meetup was described at the time of writing as "a community-led organization supported by its members, who range from students banging out code in their dorm rooms, to entrepreneurs looking to become the next Steve Jobs, to investors in search of the next big thing, to established CEOs at some of New York's top companies."[1]

We recorded and analyzed the content of 77,580 email messages (divided into 17,575 unique "threads" or discussions) sent through the Meetup's public listhost from its inception in April 2007 through April 2014. Over this seven-year period, the Meetup grew from a relatively small core of hackers, programmers, and entrepreneurs to a large and diverse collection of individuals with varying connections and interests tying them to the tech-startup community. This was the period of rapid growth when New York emerged as the second largest regional tech economy in the U.S. following Silicon Valley (e.g. Lohr, 2019). The email interactions often focus on technical subjects, and thus leave traces of the specialized knowledge possessed by individual participants. A Meetup member—for example, a novice tech entrepreneur at the early stage of founding a firm—might post an email asking for information on intellectual property law or some other topic of interest. More experienced members then respond with useful bits of information based on their own knowledge or experience.

By focusing on the content and structure of these email exchanges, we study the output of repeated—and naturally occurring—cooperative behavior consisting of the voluntary provision of knowledge and informational resources to others at a cost (i.e. of time and energy) to oneself (Baldassarri, 2015; Fowler and Christakis, 2010; Nowak, 2006; Opper and Nee, 2015; Tsvetkova and Macy, 2014). Contributors receive no direct financial compensation for their cooperative efforts. However, the cumulative pooling of bits of useful information contributes to knowledge creation and its spillover into the tech community. In effect, the email threads provide a "crowdsourcing" platform that draws on the diverse expertise of participants to create a public good—a body of information and knowledge accessible to all members (Doan et al., 2011).

Our methodology offers several advantages. By relying on behavioral traces produced without any researcher–participant interaction, we provide observations of human cooperation *in vivo* without what is classically known as the "Hawthorne effect," where the observer's presence might influence the behavior of the subjects studied (e.g. Mayo, 1993). Unlike a standard survey-based approach, our analysis is also not restricted to a fixed population; we are able to observe numerous successive cohorts of new entrants to the Tech Meetup, gather observations for every time period in which these members participated in the email discussions, and make robust comparisons both across time for individual members and across cohorts who entered in different periods. Finally, by collecting longitudinal observations of individual behavior, we are able to analyze both individual- and group-level trends concurrently without the problem of "ecological fallacy," in which the researcher attributes group-level properties to individuals within the group (e.g. Piantadosi et al., 1988). By mapping the structure of interactions between individuals with overlapping knowledge and expertise—and doing so repeatedly over numerous time periods—we distinguish individual- from group-level trends and demonstrate the evolution of both over these crucial years.

The remainder of the paper proceeds as follows. Section 2 introduces the data set of email discussions among New York Tech Meetup members. Section 3 describes the content-analytic methods we applied to induce distinct clusters of discussions from the text of the email threads. Following this, we track the extent of content-based specialization in these email discussions over the seven-year observation period using a novel measurement based on the bipartite network of email threads and user participants. Section 4 describes the statistical methods employed while Section 5 presents results from these analyses. By way of preview, we show that the

[1] Since the time of this original writing and our data collection, the non-profit New York Tech Meetup has merged with the New York Tech Council to form a new organization called the NY Tech Alliance. Its self-description on the organization's web site has changed accordingly.

typical new participant developed an increasingly specialized style of discussion participation in later periods compared to earlier periods, supporting the thesis that the tech cluster moved over time toward a clearer "division of knowledge" reflecting specialized knowledge and interests in place of generalism. Moving to the level of the whole network as opposed to individual participants, Section 6 shows that the email discussions as a whole still encompassed a diverse and cross-cutting array of content over time even as the individuals producing them came to be more specialized. In Section 7, we conclude by urging greater use of web-based repositories not just for "big" data but also for the "long" records of researcher-unobstructed social interaction they can provide.

2. The New York Tech Meetup email network

We implemented code written in Python to gather archived public records of email interactions for the NY Tech Meetup (NYTM hereafter) group (http://www.meetup.com/ny-tech/) on Meetup.com. The Meetup site provides a platform where users can create groups based on their interests and arrange in-person "meetups" for themselves and other users who join the group. As of this writing, the site boasts over 32 million members and 288,000 groups in 182 countries. The NYTM was founded in 2004 as one of the early groups on the platform. The organization presents monthly meetings where members can purchase tickets and view live demonstrations from technology-based startup companies. Perhaps more importantly, the NYTM provides a site for "networking" among professionals in the tech startup community. This function extends beyond the in-person meetings to the public email listhost maintained by the group.

The first recorded email is dated April 2, 2007. For each email in the archive, we recorded: (a) the sender ID, (b) the date and time it was sent, (c) the subject thread, and (d) the email text. The data collection garnered 79,878 emails sent over a seven-year period between April 2007 and April 2014. These emails spanned a total of 17,850 unique subject threads, meaning that the average thread featured between 4 and 5 emails (mean = 4.47). However, our analysis excludes 2,298 emails that we classified as "spam" and/or advertisements. We identified these emails by (a) hand-coding a random sample of 2,000 emails and (b) using these hand-coded examples to categorize the remaining emails using a nearest-neighbor matching algorithm, though we obtained virtually identical results when these emails were instead included in the sample. In addition to the mailing list, the NYTM also maintains a message board for use by members. We chose to focus on the mailing list because it is used far more frequently; whereas the mailing list

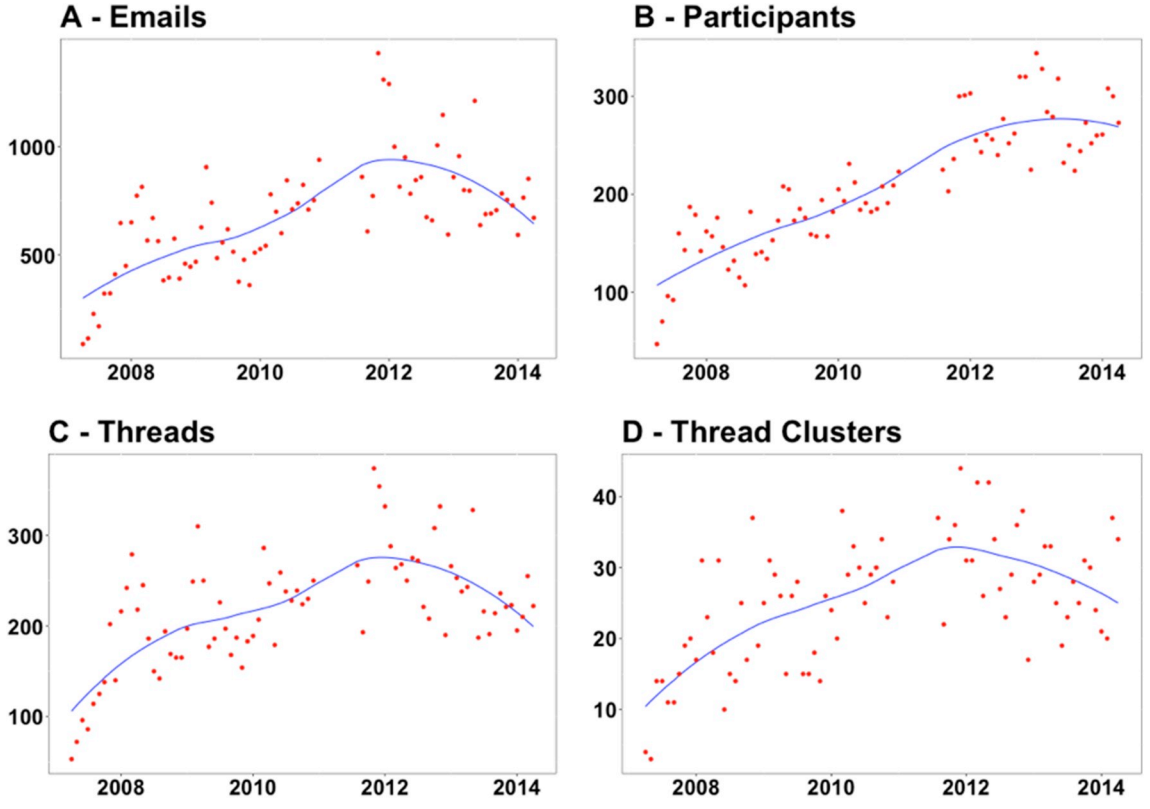

Fig. 1. Monthly growth in contributions to New York Tech Meetup email threads, 2007–2014. **Note:** Panels depict trends for (A) number of unique emails sent through the listhost, (B) number of unique users sending these emails, (C) number of unique subject threads in which emails appeared, and (D) number of unique topical clusters of email threads. Topical clusters are based on adaptive K-means clustering of the text content of the emails. Data points for each month are shown in red and the trend line is depicted in blue using local polynomial smoothing. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

averages around 30 emails per day, the message board sometimes goes unused for several days at a time.

The Meetup archive does not include any emails sent between January 1, 2011 and July 25, 2011. From examining the "bookend" emails on December 31, 2010 and July 26, 2011, we did not find any announcement indicating that the listhost had actually been taken "offline" during the intervening period. Given the wealth of data available over the rest of the seven-year period and the relative consistency of the reported time trends, we do not believe that the results would change with inclusion of the emails missing during this window. Since we have only a few days' worth of emails for July 2011, we have also excluded these from analysis, leaving us with data coverage for 79 months.

3. Content analysis of email threads

Our main unit for text-based analysis is the subject thread. Accordingly, we append the "bag of words" appearing in each particular email to the list of words appearing in other emails from the same thread. We take this approach for several reasons. First, the subject thread—as the term implies—lends organization to the emails appearing within it. The title of the thread denotes the topic of discussion, and the subsequent emails appearing in the thread follow from this topic. So, it is natural to view the emails appearing within the same subject thread as being contributions to the same discussion.

However, there are also practical reasons for organizing the data this way. When applying an automated algorithm to the collection of the email text, it is sometimes impossible to distinguish the "new" text contributed by an email from the trailing text contributed by previous emails in the same subject thread. Heuristics that could be applied to get rid of the trailing text often had the unintended consequence of "lopping off" the new text when applied to a different set of emails generated from a different user email platform, thereby causing loss of data. This problem was especially difficult because, unlike most other features of the Meetup site, the NYTM mailing list archive is not searchable through the site's Application Programming Interface (API) and therefore had to be gathered using "web-scraping" techniques.

Our solution was to retain both the new and trailing text for each email, then combine emails from the same thread into one document for textual analysis. To ensure that the trailing text repeated from previous emails did not artificially inflate word frequency counts, we also restricted the cells in the word-by-document matrix to have values of either 1 (indicating that word *i* appeared in thread *j*) or 0 (indicating the absence of word *i* in thread *j*). Since we use threads rather than individual emails as the key units, we also treat each email as occurring on the day that the first email in its subject thread was sent. This simply means that a single subject thread can only influence the results for one month.

Fig. 1 tracks the growth over time in online New York Tech Meetup-related activity. Email activity through the NYTM listhost increased dramatically over the observation period in terms of the number of unique emails sent to the listhost (panel A), the number of unique participants who sent these emails (panel B), and the number of distinct subject threads comprising these email chains (panel C).

We next analyzed the content of the email threads with the goal of identifying the different specializations represented in the email discussions for each month in our seven-year observation period. To this end, we used the words present in any given thread as a simple but robust indicator of its substantive content (e.g. Rule et al., 2015). This inductive approach allows the specialized content of each email thread to emerge organically from the participants. The "bag of words" for each subject thread *j* in month *t* was read into the *R* software and analyzed using the "tm" package (Feinerer and Hornik 2018). The document for each thread was rid of punctuation, numbers, empty whitespace, and all words were transformed to lowercase. Next, we removed a standard list of English "stopwords" that appear too frequently to be of use in distinguishing one document from another. To further clean the text, we also compared each word in each document against the dictionary contained in the "hunspell" package in *R* and only retained words that found a correct match (Ooms, 2018). Finally, we "stemmed" the words using Porter's algorithm (Porter, 1980).

The cleaned text for all threads in month *t* is transformed into a term-document matrix. We pared down this matrix to only include terms that appeared in at least one percent of documents, removing especially sparse terms that appeared in just one or exceedingly few email threads. The remaining cells are weighted using *term frequency – inverse document frequency* (TF-IDF), formally defined as

$$\frac{b(u,\ d)}{\sum_k n(k,\ d)} \times \log \frac{n(D)}{n(u,\ D)} \tag{1}$$

where *b(u,d)* indicates the presence (1) or absence (0) of term *u* in document *d*; *n(k,d)* indicates the number of terms in *d*; *n(D)* is the total number of documents; and *n(u,D)* is the number of documents in *D* containing term *u* (Salton and Buckley, 1988). The first fraction weights the appearance of term *u* relative to the opportunities for its appearance in that particular document (e.g. based on the length of the document and number of terms used), while the second fraction indexes the "distinctiveness" of term *u* based on the regularity with which it appears in the documents.

The next step is to cluster documents on the basis of semantic similarity. This is how we identify the *thread clusters* present for a given month, each of which consists of subject threads with similar text content. We first transform the weighted term-document matrix into a document-by-document cosine distance matrix. For each month, we then apply an adaptive K-means clustering algorithm to the distance matrix using the "akmeans" package in R (Kwak, 2014). Beginning from two clusters, this algorithm successively divides clusters, testing after each division whether the new set of clusters satisfies a threshold condition. Once all clusters meet the threshold condition, the algorithm stops producing additional clusters. For any given month, then, the number of clusters can theoretically vary from 2 to *N*, where *N* is the total number of email threads retained for analysis.

After exploratory analyses, we visually inspected the clustering results for 12 selected months evenly spaced over the seven-year

observation period. By examining influential terms with high *term frequency – inverse document frequency* (TF-IDF) scores, we tried to remove potential "noise" from the data: email form words, names, email signatures, and salutations. After removing these words and re-running the analysis, we found similar results. The results presented are from the cleaned data. While we cannot dismiss the possibility of other undetected noise in the data, as is nearly always the case when working with "messy" user-produced text, we were encouraged by the clarity and consistency of the observed patterns. After applying this process of data clustering and cleaning to all months in the observation period, panel D of Fig. 1 shows that—as with our previous descriptive measures of email thread participation—we observe growth over time in the number of distinct specialized discussions taking place during a typical month.

4. Statistical approach to modeling individual-level specialization

Getting closer to the micro-level, our main analysis focuses on trends in individual participation in the increasingly large and diverse number of topical areas represented in the email threads. One possibility is that in a rapidly growing tech economy, new people entering the ecosystem have increasingly broad and diverse interests. A second possibility is that as the market for knowledge workers thickens, individuals become increasingly specialized while the number of distinct specializations available to new entrants increases simultaneously. This second trend may reflect an increasing market demand for *specialists* with knowledge and interests that focus on one particular "niche" of knowledge and capability.

To adjudicate between these patterns, we develop a scale to measure *specialization* versus *generalism* among individual participants using a Herfindahl concentration index (Rhoades, 1993). Economists typically use this index to measure the competitiveness of markets. If one or a few firms dominate with huge market shares, the market is not very competitive. In our analysis, the thread clusters are "firms" that compete for the attention of participants. If nearly all of a participant's contributions to discussion occur within one or a few topical clusters, then we can classify this participant as relatively specialized. Conversely, participants who spread their contributions more evenly across topical clusters are generalist. We compute this concentration index separately for each participant in each month that he or she participated in the email threads, while also adjusting for the total number of emails he or she sent and the varying size of the topical areas. Fig. 2 uses a toy example to illustrate how we distinguish specialists from generalists using this approach.

More formally, for each person i who contributed to the email discussions for month t, we calculate

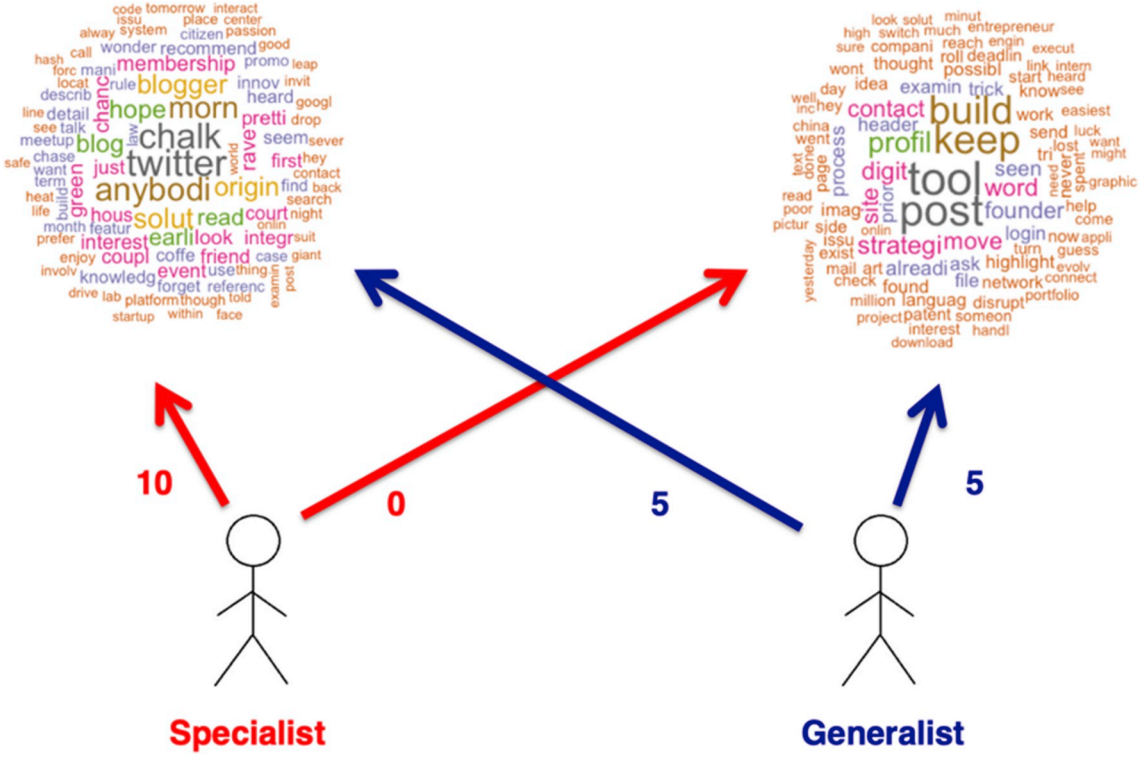

Fig. 2. Toy illustration of specialists and generalists in email discussions. **Note**: Figure depicts two hypothetical users contributing emails to two hypothetical clusters of email threads. The specialist sends all ten emails to threads in the same cluster, exhibiting a maximally high level of specialization by our measurement. The generalist divides her emails equally across the two clusters, exhibiting a tendency toward generalism. The clusters are represented as word clouds with stemmed key terms.

$$C_{it} = \sum_{j=1}^{K} s_{ijt}^2 \tag{2}$$

where s_{ijt} is the proportion of i's emails occurring in threads belonging to cluster j and K is the total number of thread clusters for the month. However, this raw concentration score for each individual will depend not just on that person's relative concentration of email activity *within* rather than *across* clusters, but also on the relative size of the clusters observed for that month. Consider the simple case of one month featuring email clusters of relatively equal size and another month featuring one large cluster and several small ones. The C_{it} score will be systematically higher in the second month—the one with unequally sized clusters—simply because an email sent out at random will be more likely to fall into the large cluster for no other reason than the cluster's size relative to others.

To correct for this size bias, we can plot the data as a bipartite (two-mode) network of *people* tied to *email threads* through the linkage created by sending an email to that thread. For each month, we generate 1,000 randomly rewired permutations of this bipartite network containing the same people, email threads, and emails as the observed network, but in which all of the ties between people and threads occur randomly. For this procedure, we used the "bipartite" package in R (Dormann et al., 2008). Then, we re-compute C_{it} for each person in each simulated network and take the average score for person i across all 1,000 rewired networks as an indicator of that person's "expected" score under conditions of random email activity. The adjusted specialization score for person i in month t is simply

$$C_{it}^{*} = C_{it} - E(C_{it}) \tag{3}$$

where $E(C_{it})$ is the expected score. Since both C_{it} and $E(C_{it})$ can theoretically range from 0 to 1, the adjusted score can range from −1 to 1. In practice, the scores heavily concentrate on the unit interval. The final dataset contains 3,989 unique persons and 16,363 person-month observations, meaning that the average person contributes about 4.10 data points by appearing in the email network for multiple months.

After computing specialization scores for each person-month observation, we estimate a linear mixed-effects regression model (Gelman and Hill, 2007) taking the form

$$Y_{it} = \alpha_{i,c} + \beta X_{it} + \sum_{m=1}^{N} \theta_m V_{it,m} + e_{it} \tag{4}$$

where $\alpha_{i,c}$ is a varying intercept that is particular to each person i and entry cohort c (the month that the person first enters the data set). Substantively, this means that we estimate a multilevel model in which person-month observations are nested within persons who are in turn nested within entry cohorts. Elsewhere in the model, X_{it} gives the time elapsed since i's first appearance in the data (scaled in terms of years) and β gives the regression coefficient for the expected increase or decrease in specialization accruing from an additional year since first entry. To ensure meaningful comparisons, we only compare people with the same volume of email activity. Thus, $V_{it,m}$ is a binary indicator of whether person i sent m emails in month t, where N is the maximum observed email volume and θ_m is the estimated fixed effect on specialization of sending m emails (we use the median as a reference category). Finally, e_{it} is a residual term capturing additional variation in the outcome variable.

Using this model, we focus on two patterns. The first (reported verbally in the next section) concerns the presence or absence of a trend toward increasing specialization over time *within persons*. Substantively, this effect is captured by β in equation (4). The second pattern concerns whether there is increasing individual specialization over time *across successive entry cohorts*. These data points are found using portions of the varying intercept, which is estimated more formally as

$$\alpha_{i,c} = \gamma + \sigma_c + \rho_{i,c} \tag{5}$$

where γ is the overall average specialization for a new entrant to the data set (X_{it}=0) with median-level email volume ($V_{it,m}$= 1 for median m); σ_c is the varying portion of this intercept that is particular to entry cohort c; and $\rho_{i,c}$ is the varying portion of the intercept that is particular to person i within cohort c. To obtain the expected average specialization for the members of each entry cohort c, we sum σ_c and γ, leaving out the additional person-specific variation within each cohort.

5. Trends in specialization over time

In the statistical analysis, each participant's contributions in the email discussions are compared both to their behavior from earlier time periods and to other participants who entered the discussion network either earlier or later. Using a mixed-effects panel regression model, we estimated two trends. First, as regards the *within-person* individual trend toward either increasing or decreasing specialization among people who participate across multiple time periods, we did not find a statistically significant linear trend in either direction ($B = 0.001$ for each year since first appearing in the data; S.E. $= 0.001$, $P = .317$). Using the same model, secondly, we estimated the *between-cohort* individual trend toward increasing or decreasing average specialization across successive groups of new entrants. As shown visually in panel A of Fig. 3, this result strongly supports the hypothesis that newer cohorts exhibited greater specialization than older cohorts. While Fig. 3 displays a nonparametric smoothed fit of the time trend, the corresponding linear best fit of the same time trend has an R^2 of 0.42 and is significant at the $P < .001$ level.

These patterns suggest that the types of informational resources contained in this public good evolve over time to match the growth

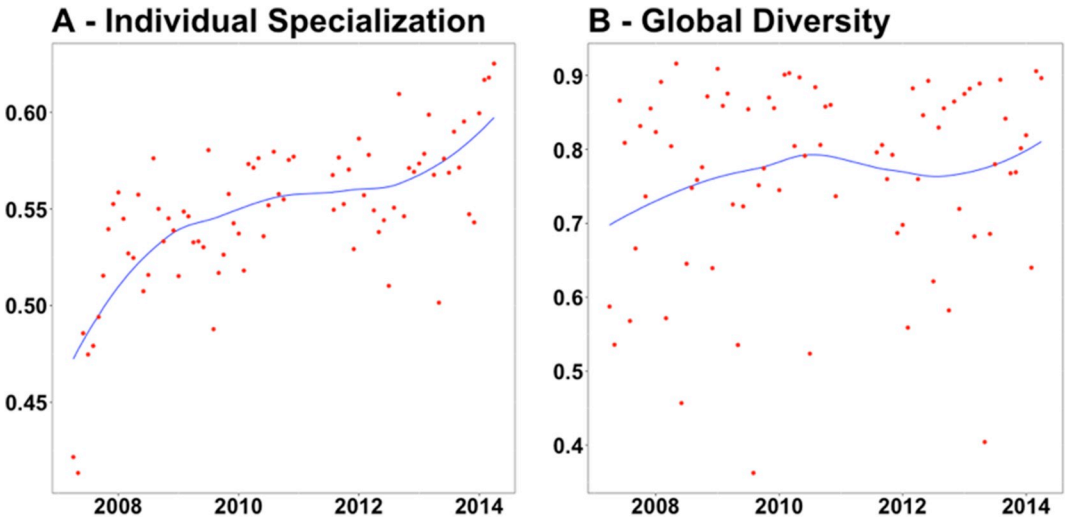

Fig. 3. Individual specialization and global diversity over time. **Note**: Panel A is based on individual users' specialization scores for each month, which measure the concentration of a person's emails within thread clusters, minus the expected concentration if that person had sent the same number of emails but distributed them randomly across clusters. Typical point estimates for each month's cohort are based on a mixed-effects statistical model that adjusts for time in the group (which is set here to zero) and each person's volume of email activity during the month (which we hold constant at the median value for this visualization). Panel B is based on global or network-level diversity of content, which is defined as $1 - \sum_{j=1}^{K} s_j^2$ where s_j is the proportion of total emails occurring in threads pertaining to cluster j and K is the number of unique clusters. Data points for each month are shown in red and the trend line is depicted in blue using local polynomial smoothing. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

and complexity of the emerging industrial cluster. Email threads in earlier years feature a relatively undifferentiated mass of discussion on topics of broad interest to the nascent "technologist" community. Over time, these discussions become increasingly specialized, as more (and different) types of participants become involved. Interpreting these patterns, we suggest that earlier discussions on topics of broad interest serve to solidify the group's emerging identity and demonstrate the potential for productive cooperative exchange. With this foundation, group members later venture into increasingly specialized topics where more knowledge and information is needed to serve emerging "niches."

6. Trends in network-level content diversity

A natural next question is whether increasing individual specialization also implies an increasingly narrow focus at the group-level. Put differently, does a population of increasingly specialized participants lead to an increasingly diverse or narrow pool of available information? Imagine, for example, if one discussion drew 90 percent of the email activity while thirty others shared the remaining 10 percent. While the number of specializations represented in the email traffic would be high, the heavy concentration of that activity within one discussion suggests a lack of diverse knowledge (e.g. Evans, 2008). To assess the extent of diversity, we estimated the global Herfindahl concentration index for each month in the observation period based on each topical cluster's share of emails contributed. For a given month, we measure the network-level diversity of content as one minus this Herfindahl concentration score; therefore, a diversity score of 1 would indicate a lack of concentration and an even distribution of email contributions across clusters, while a score of 0 would indicate a maximally concentrated distribution. Panel B of Fig. 3 shows that overall contributions moved over time toward slightly greater diversity even as the number of topical areas increased and individual participants became increasingly specialized. Rather than opposing one another, individual specialization and group-level diversity appear to correlate together rather closely.

As specialization proceeds apace, however, does this also lead to balkanization in which distinct communities of communicating specialists become sealed off from one another? To answer this question, we further analyzed the network of co-participation in email threads for each month, where two email threads sharing overlap in participation are "linked" in the network. The data points for this analysis are based on *modularity*, frequently applied in network analysis to identify "communities" or groups of objects that are densely connected to one another and less so to objects in other groups (Newman and Girvan, 2004). For each month, we begin with the bipartite network of persons tied to email threads in which they participated. Then, we take the transpose of the person-by-thread matrix to obtain a weighted network of email threads linked through numbers of co-participants. We next label each thread according to the content cluster assigned by the K-means clustering algorithm. Finally, we compute the modularity score for the thread co-participation network.

This score has a simple substantive interpretation: the proportion of network ties (or edges) occurring *within* groups compared to the proportion we would expect by random chance. Again, the purpose of comparing the observed network to a hypothetical random or "null" configuration with the same size and degree distribution is to ensure measurements that are independent of changes month-

to-month in the size of the network. More formally,

$$Q = \sum_i \left(e_{ii} - a_i^2\right) \tag{6}$$

where groups are indexed by i, e_{ii} is the proportion of network ties occurring within group i, and $a_i = \sum_j e_{ij}$ where other groups are indexed by j (Newman and Girvan, 2004).

Panel A of Fig. 4 displays this network evolution visually, illustrating the previously uncovered pattern of a growing and increasingly diverse collection of topical areas in the email threads.[2] More formally, we were interested in the degree of *cross-pollination* between threads pertaining to different topical areas. To assess this pattern, as described above, we track the trend in network *modularity,* defined again as the proportion of co-participation links occurring *within* clusters of similarly categorized email threads minus the proportion we would expect under assumptions of random sorting (Newman and Girvan, 2004). By this metric, panel B of Fig. 4 suggests a relatively stable level of cross-pollination—or interaction across topical boundaries—across the observation period, as indicated by the consistently low modularity of the co-participation network. While individuals are becoming increasingly specialized, in other words, there is still "bridging" interaction across topical boundaries rather than balkanization.

As a further test of trends in the thread co-participation network, we also computed *participation coefficients* for all nodes (email threads) across all months in the observation period. The participation coefficient indexes the extent to which the network connections or edges (in this case, co-participation across threads) incident to a given node extend beyond that node's module (in this case, topical clusters). More formally, Guimera and Amaral (2005a) define the participation coefficient for a given node i as

$$P_i = 1 - \sum_{s=1}^{N_M} \left(\frac{k_{is}}{k_i}\right)^2 \tag{7}$$

where modules are indexed by s, N_M is the total number of distinct modules in the network, k_{is} is node i's number of links to nodes in module s, and k_i is i's total degree (also see Guimera and Amaral, 2005b and Baggio et al., 2015; for the relevant *R* package, see Watson, 2018). In our application, an email thread having a participation coefficient of 1 would mean that the co-participation links involving that email thread were evenly distributed across all topical clusters; a coefficient of 0 would indicate that co-participation links involving that thread fell entirely within its own topical cluster. Panel C in Fig. 4 shows that the average participation coefficient in the thread co-participation network increased over the observation period. By this metric, cross-pollination across topical clusters actually became relatively more common over time compared to purely within-cluster discussions.

7. Conclusion

As distinct from economies that rely on physical inputs or natural resources, innovative activity in the rising knowledge-based economy depends critically on knowledge-sharing and spillover both within and across the boundaries of individual firms. Specialized knowledge forms the essential component of knowledge spillovers—the positive externalities that follow when specialists with varying bits of knowledge are put in a position to interact and learn from one another (Storper and Venables, 2004). Such spillovers are in turn linked to the agglomeration of skills and human capital that drives industrial growth (e.g. Glaeser, 1999).

This article has documented the structure and emergence of an ecology of specialized knowledge in one contemporary innovation cluster. At the outset, we could have imagined this ecology conforming to either of two "ideal types." In one version, the division of knowledge could hinge on Elon Musk-like knowledge entrepreneurs—generalists capable of carrying insight across multiple domains. In the second version—closer to the pattern we uncover—the division of knowledge hinges on distinct but overlapping communities of specialists. Akin to parallel processing in computing, a complex task that might never be completed by a single computer can become trivially easy when distributed among many, with each given only a small portion of the total workload. We might term this phenomenon *parallel knowledge processing* wherein the cooperation of knowledge workers within many specialized networks and groups contributes to the broader division of labor at the level of the industrial cluster. In practice, these two versions of knowledge organization are likely co-dependent. Elon Musk needs the parallel knowledge processing capacity of Silicon Valley in order to be Elon Musk. In Silicon Valley and other innovation hubs, in other words, generalist entrepreneurs rely on the parallel knowledge processing of many specialists for entrepreneurial success.

Consistent with an emergent trend toward the industrial cluster organized as parallel processing, individual participants in the New York Tech Meetup became more specialized over time, increasingly concentrating their activities within email threads with similar content. Yet, at the same time, the email discussions as a whole maintained a relatively high level of diversity in content. This is consistent with classic accounts of the division of labor, as when a manufacturing firm making several distinct products creates a multidivisional structure allowing different branches of workers to focus on producing just one of these products with maximal efficiency. In knowledge- and technology-based economic sectors where the boundaries of firms are increasingly blurred (Owen-Smith and Powell, 2004; Saxenian, 1994), we can analyze and uncover a similar division of labor based on the organization of specialized knowledge and using people rather than firms as the key unit of analysis.

To be sure, there are limitations to this analysis. Relying on behavioral traces of public discussions still provides a limited image of

[2] These network visualizations were produced using the "igraph" package in R (Csardi and Nepusz, 2006).

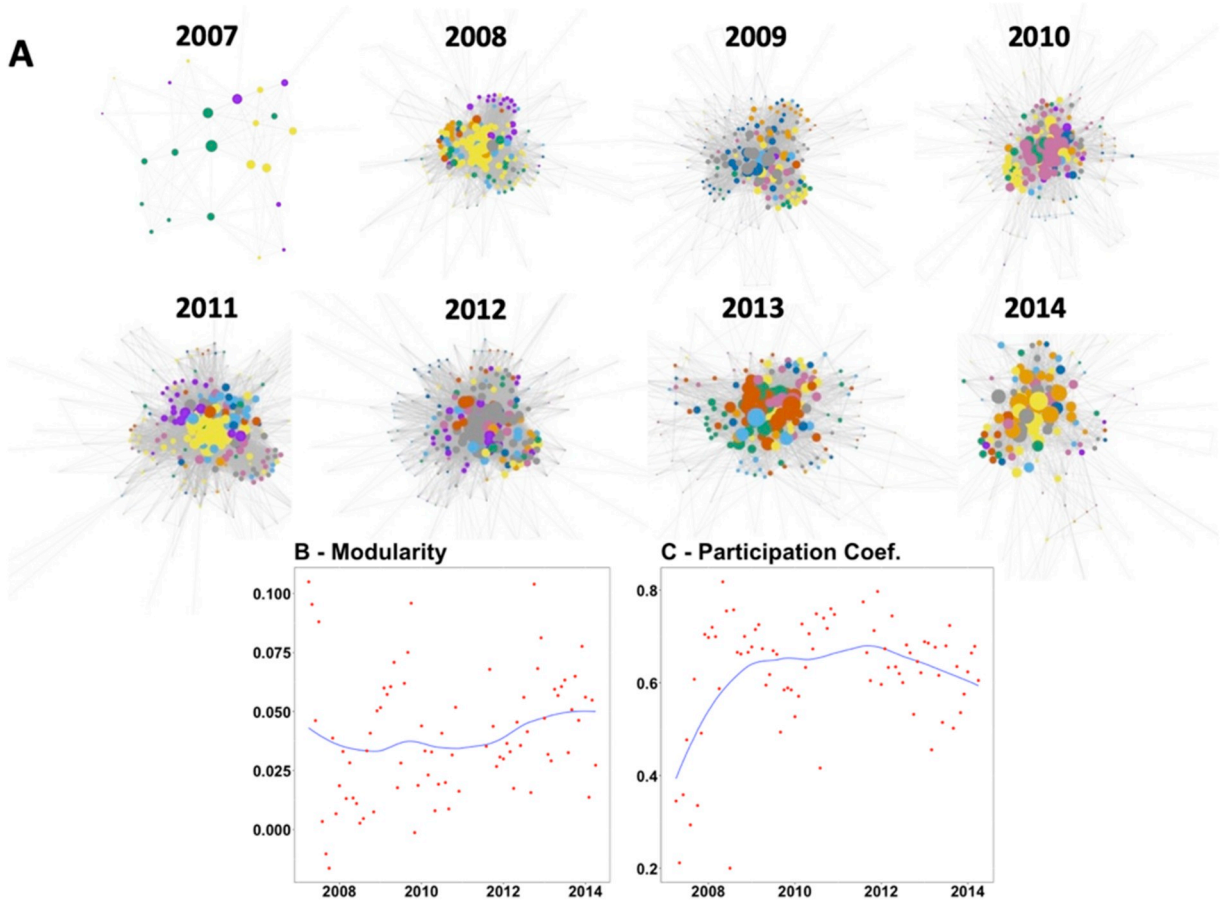

Fig. 4. Networks of co-participation across thread clusters over time. **Note**: In Panel A, network visualizations are plotted based on data from April of each year. Since the email list was offline during April 2011, we use the next available month for this year, which was August. Nodes corresponding to each email thread are sized by degree and colored according to cluster membership. The edges in the network are generated by co-participation in which the same users contributed emails to multiple threads. Panels B and C show time trends in which data points for each month are shown in red and the trend line is depicted in blue using local polynomial smoothing. Panel B displays network modularity, while Panel C shows the mean participation coefficient for the network. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

the knowledge and information carried by individual participants. Yet other benefits speak to the potential richness of analyses drawing on new and emerging sources of digitally recorded social-behavioral data across a wide variety of domains (Golder and Macy, 2014; Kossinets and Watts, 2006; Lazer et al., 2009; Salganik, 2018). One salient challenge in previous work using such data has been relating the online patterns to offline contexts. Our goal was first to use a long-term archive of "digital traces" generated by participation in a group of tech entrepreneurs and knowledge workers to detect offline changes in this emerging tech cluster in a major American city. Our approach suggests attention to web-based repositories as a source of "big" (i.e. large sample) data, but also as an emerging source of "long" data. An advantage of these targeted long-term databases is that they can be used to study historical processes without having to rely on *post hoc* accounts. When they track individual behavior over time, they also allow us to simultaneously measure changes at both individual- and group-levels. This distinction carries special importance when groups are in constant flux due to the "churning" of entry and exit and the link between "micro" and "macro" behaviors is not straightforward (Schelling, 1978). Since countless social situations have this quality of being continuously changed or reproduced by individual sorting, we expect that the approach taken here could find general application in numerous other research contexts.

Acknowledgment

For their helpful comments on earlier versions of this article, we wish to thank Michael Macy as well as members of the Economy and Society Lab at Cornell University. We are also grateful for helpful research assistance from Sirui Wang and Sohyeon Hwang. The second author wishes to also acknowledge grant funding from the Marianne and Marcus Wallenberg Foundation and from the John Templeton Foundation.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ssresearch.2019.102377.

References

Baggio, J.A., Brown, K., Hellebrandt, D., 2015. Boundary concept or bridging concept: a citation network analysis of resilience. Ecol. Soc. 20, 2.
Baldassarri, Delia, 2015. Cooperative networks: altruism, group solidarity, reciprocity, and sanctioning in Ugandan producer organizations. Am. J. Sociol. 121, 355–395.
Bell, Daniel, 1973. The Coming of Post-Industrial Society. Basic Books, New York, NY.
Castells, M., 1996. The Rise of the Network Society. Blackwell Publishers, Oxford.
Cometto, Maria Teresa, Piol, Alessandro, 2013. Tech and the City: the Making of New York's Startup Community. Mirandola, New York, NY.
Csardi, G., Nepusz, T., 2006. The igraph software package for complex network research. InterJ. Complex Syst. 1695. http://igraph.org.
Doan, Anhai, Ramakrishnan, Raghu, Halevy, Alon Y., 2011. Crowdsourcing systems on the world wide web. Commun. ACM 54, 86–96.
Dormann, C.F., Gruber, B., Fruend, J., 2008. Introducing the bipartite package: analysing ecological networks. R. News 8, 8–11.
Durkheim, E., 1933. The Division of Labor in Society. Free Press, New York.
Evans, James A., 2008. Electronic publication and the narrowing of science and scholarship. Science 321, 395–399.
Feinerer, Ingo, Hornik, Kurt, 2018. Tm: Text Mining Package. R package version 0.7-6. https://CRAN.R-project.org/package=tm.
Fowler, James H., Christakis, Nicholas A., 2010. Cooperative behavior cascades in human social networks. Proc. Natl. Acad. Sci. 107, 5334–5338.
Gelman, Andrew, Hill, Jennifer, 2007. Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press, Cambridge.
Glaeser, Edward L., 1999. Learning in cities. J. Urban Econ. 46, 254–277.
Golder, Scott A., Macy, Michael W., 2014. Digital footprints: opportunities and challenges for online social research. Annu. Rev. Sociol. 40, 129–152.
Grant, Robert M., 1996. Toward a knowledge-based theory of the firm. Strateg. Manag. J. 17, 109–122.
Guimera, Roger, Amaral, Luis A. Nunes, 2005. Functional cartography of complex metabolic networks. Nature 433, 895–900.
Guimera, Roger, Amaral, Luis A. Nunes, 2005. Cartography of complex networks: modules and universal roles. J. Stat. Mech. Theory Exp., P02001
Hidalgo, César A., 2015. Why Information Grows. Basic Books, New York, NY.
Jackson, Matthew O., 2008. Social and Economic Networks. Princeton University Press, Princeton, NJ.
Kossinets, Gueorgi, Watts, Duncan J., 2006. Empirical analysis of an evolving social network. Science 311, 88–90.
Kwak, Jungsuk, 2014. Akmeans: Adaptive K-Means Algorithm Based on Threshold. R package version 1.1. https://CRAN.R-project.org/package=akmeans.
Lazer, David, et al., 2009. Life in the network: the coming age of computational social science. Science 323, 721–723.
Lohr, Steve, 2019. It Started with a Jolt: How New York Became a Tech Town. *New York Times*. https://www.nytimes.com/2019/02/22/technology/nyc-tech-startups.html. (Accessed 23 September 2019).
Mani, Dalhia, Moody, James, 2014. Moving beyond stylized economic network models: the hybrid world of the Indian firm ownership network. Am. J. Sociol. 119, 1629–1669.
Mayo, Elton, 1993. The Human Problems of an Industrial Civilization. MacMillan, New York, NY.
Mokyr, Joel, 2002. The Gifts of Athena: Historical Origins of the Knowledge Economy. Princeton University Press, Princeton, NJ.
Newman, M.E.J., Girvan, M., 2004. Finding and evaluating community structure in networks. Phys. Rev. E 69, 026113.
Nowak, Martin A., 2006. Five rules for the evolution of cooperation. Science 314, 1560–1563.
Ooms, Jeroen, 2018. *Hunspell: High-Performance Stemmer, Tokenizer, and Spell Checker.* R Package Version 3.0. https://CRAN.R-project.org/package=hunspell.
Opper, Sonja, Nee, Victor, 2015. Network effects, cooperation and entrepreneurial innovation in China. Asian Bus. Manag. 14, 283–302.
Owen-Smith, Jason, Powell, Walter W., 2004. Knowledge networks as channels and conduits: the effects of spillovers in the boston biotechnology community. Organ. Sci. 15, 5–21.
Padgett, John F., Powell, Walter W., 2012. The Emergence of Organizations and Markets. Princeton University Press, Princeton, NJ.
Piantadosi, Steven, Byar, David P., Green, Sylvan B., 1988. The ecological fallacy. Am. J. Epidemiol. 127, 893–904.
Porter, M.F., 1980. An algorithm for suffix stripping. Program 14, 130–137.
Powell, Walter W., Snellman, Kaisa, 2004. The knowledge economy. Annu. Rev. Sociol. 30, 199–220.
Rhoades, Stephen A., 1993. The Herfindahl-Hirschman index. Fed. Reserve Bull. 79, 188–189.
Rule, Alix, Cointet, Jean-Philippe, Bearman, Peter S., 2015. Lexical shifts, substantive changes, and continuity in state of the union discourse, 1790-2014. Proc. Natl. Acad. Sci. 112, 10837–10844.
Salganik, Matthew J., 2018. Bit by Bit: Social Research in the Digital Age. Princeton University Press, Princeton, NJ.
Salton, Gerard, Buckley, Christopher, 1988. Term-weighting approaches in automatic text retrieval. Inf. Process. Manag. 24, 513–523.
Saxenian, AnnaLee, 1994. Regional Advantage: Culture and Competition in Silicon Valley and Route 128. Harvard University Press, Cambridge, MA.
Schelling, Thomas C., 1978. Micromotives and Macrobehavior. Norton, New York, NY.
Simon, H.A., 1981. The Sciences of the Artificial. MIT Press, Cambridge, MA.
Smith, A., 1776. The Wealth of Nations. University of Chicago Press, Chicago, IL.
Storper, Michael, Venables, Anthony J., 2004. Buzz: face-to-face contact and the urban economy. J. Econ. Geogr. 4, 351–370.
Tsvetkova, Milena, Macy, Michael W., 2014. The social contagion of generosity. PLoS One 9, e87275.
Varian, Hal R., 1995. The information economy. Sci. Am. 273, 200–201.
Watson, Christopher G., 2018. BrainGraph: Graph Theory Analysis of Brain MRI Data. R package version 2.2.0. https://CRAN.R-project.org/package=brainGraph.
Whittington, Kjersten Bunker, Owen-Smith, Jason, Powell, Walter W., 2009. Networks, propinquity, and innovation in knowledge-intensive industries. Adm. Sci. Q. 54, 90–122.