Measuring Mohr Social Capital

Monica Lee[a,*]

Amaç Herdağdelen[a]

Minsu Park[b]

John Levi Martin[c]

* To whom correspondence should be addressed; [a] Facebook Inc.; [b] New York University Abu Dhabi; [c] University of Chicago.

Measuring Mohr Social Capital

Abstract

We here bring together two different traditions of thinking about social capital. One, the *Tocquevillian,* looks to associations and group memberships as the core of social capital. The other, the *Colemanian*, looks to interpersonal networks as the core of social capital. We argue that the most common way of articulating how humans use these types of relationships in different ways—the distinction between "bridging" and "bonding" social capital—is epistemically unstable. What might be possible, however, is to use the insights developed by Ronald Burt regarding tie *non-redundancy* to study associational social capital. We do this by drawing on the insights of the approach consistently adopted and developed by John Mohr, which emphasizes *duality* and *diversity*, to develop measures of group affiliation-based social capital. We accordingly, for both Tocquevillian and Colemanian social capital, distinguish measures that focus on the *mass* of social capital from those that focus on its *diversity.* To illustrate, we assess the degree of social capital of all resulting types for 77 Million U.S. Facebook users who are active in Facebook Groups, showing that our understanding of who has the most social capital varies greatly by whether we are considering Tocquevillian or Colemanian capital, and whether we are focusing on mass or diversity.

1. Introduction and Overview

There have, as we go on to show, historically been two different ways of understanding social capital. One sees this as fundamentally about the presence of robust voluntary associations, and the other sees it as patterns of interpersonal connections. The intersection between these—the capacity of social groupings to scaffold new forms of interpersonal relations—has not been explored, even though this relation seems implied by some of the core orienting conceptions of mathematical sociology, that of duality. Following Breiger (1974) and Breiger and Mohr (2004), we use the duality inherent in a persons × groups matrix to understand the *implicit* social ties established by groups. This conception can, we argue, better reach some insights about the nature of social capital than existing ways of trying to partition between "bridging" and "bonding" capital. Building on the difference between social capital seen as group memberships, and social capital seen as interpersonal ties, we begin by reviewing theories of social capital, point to a paradox in some current conceptions, and then lay out our own approach. For both social capital as group memberships (which we call "Tocquevillian") and social capital as individual relations (which we call "Colemanian"), we distinguish measures that focus on the total *mass* of capital from those that focus on its *diversity*. We illustrate these with data on American adults' participation in Facebook groups.

2. Social Capital as Group Affiliations and Relationships

2.1 The Two Theories of Social Capital and the Two Varieties of Each

The notion of social capital—meaning the advantages coming from a stock of social relations and involvements, as opposed to the socialized form of economic capital—has been a central part of sociology and economics, yet we still find theorists struggling to free themselves from misleading assumptions associated with the term. In particular, some confusion has resulted from the fact that there are two core visions of what we mean by "social capital," which carry different connotations and direct us towards different types of quantification.

On the one hand, there is a version of social capital that goes back to de Tocqueville (1962 [1835]), and was later seized upon by those enthusiastic for theorizing the nature of American civil society (e.g., Bellah et al. 1985). In this tradition, we turn to social capital to answer the question: why did some European societies collapse into fascism while the United States remained a democracy? Mass society theorists (e.g., Kornhauser 1959) emphasized the importance of intermediary organizations in the preservation of democracy. The United States, possessing a political culture that turned on volunteerism as opposed to state intervention (Clemens 2020), seemed unusually rich in such associations. In this vision, we at least *start* from the perspective that social capital is an inherently *collective* good.

On the other hand, there is a version of social capital as an *individual* characteristic, a usage that was first kicked around informally (e.g., by Durkheim's disciple Célestin Bouglé [1926 / 1922: 50; see 43 for "intellectual capital"]), but revived in the United States in the 1980s (Coleman 1988). Although perhaps the most famous use of this approach appears to focus on a joint form of social capital (intergenerational closure—when parents know their kids' friends' parents [e.g., Carbonaro 1998]), Coleman's whole approach was rooted in the self-interest of the individual actor (Coleman 1990). It was not, therefore, the sort of understanding of social capital that

defined it as an unalloyed collective benefit, even if it was not strictly zero-sum.

Thinking of such individual-level social capital—the potential benefit in having a stock of relations—we have little difficulty anticipating a negative side of social capital. Indeed, Pierre Bourdieu (e.g., 1986) used the term "social capital" to refer to substantively very similar patterns as our "Colemanian" capital, but he interpreted the capital itself in a relational sense (that is, social capital is not merely about relations to others, but relations between *my* relations and *your* relations—I have more social capital than you if my friends are better placed than yours). In this light, social capital, like other forms of capital, is a latently antagonistic relation: there is no use in having social capital if all have it in just the same form. And indeed, empirical work, such as Beyerlein and Hipp (2005), confirmed the reasonableness of such reservations, finding that the benefits of social capital could indeed be zero-sum—my social capital comes at *your* expense. (One may think of the much-vaunted teenagers who used their internet prowess to help their neighbors get COVID-19 vaccine appointments early in vaccine rollout—denying these slots to those without connections.) Still, one could propose that this is to some extent true of all resources that can be employed in a competitive system. We will return to this issue below, but first consider the way in which similar doubts began to trouble those in the Toquevillian tradition.

The downside of the associational view of social capital comes in two forms. First, when one thinks clearly about voluntary groups, one is forced to realize that these do not only include the PTA and Sierra Club, but also Hitler's Brownshirts if not also the Mafia. It matters what the groups are trying to do. The same is true even for more generalized collective measures: Messner et al. (2004) demonstrated that while some purported measures of social capital went along with lower homicide rates, others predicted *increased* murders. Findings like this led some to admit that there could be a "dark side" to social capital—when it was used to do things of which the writer in question disapproved.[1]

But second, the very social cohesion that gives a neighborhood social cohesion when it comes to taking care of insiders can be used against outsiders. It clearly is not a recipe for civil flourishing to have the polity divide up into dense cliques in which one loves one's neighbor as oneself—and hates and fears all others. For this reason, social capital theorists have increasingly accepted Gittell and Vidal's (1998) distinction between *bonding* and *bridging* social capital. The first indicates the sort of dense web of connections that might allow for successful joint endeavors that could not be carried out by actors who were not connected by multiple ties (e.g., Greif 1989), the sort of structure that had been theorized by Granovetter (1985) under the rubric of "embeddedness." The second indicates ties that link one such dense group to another, the sort of structure that had been theorized by Granovetter (1973) in his work on "weak ties" (a contribution which was actually more about *structure* than *strength*).

This would seem to suggest that any assessment of the positive side of social capital must either look for bridging capital, or at least both bridging *and* bonding capital—bonding capital by itself is dangerous. Yet, as we go on to show, this distinction is fundamentally unstable.

---

[1]. This blatantly subjective nature of the definition was quite reasonable in the context in which Gargiulo and Benassi (1999) first used the notion of the dark side—it was about the ways that social capital could prove problematic for a *manager* attempting to "get ahead."

2.2 Duality and Bridging Capital

To explicate our claim regarding the notion of the formal instability of bridging vs. bonding capital, we draw on the Simmelian notions of duality of person and group used by Breiger (1974) and inspirational to Mohr. Following Breiger's classic work on persons and groups (1974, also see 2000), Mohr (2000; for examples, see Mohr and Duquenne 1997; Mohr and Friedland 2008; Breiger and Mohr 2004) proposed that the core principle of duality was fundamental for sociological theorizing. This approach can also be used to clarify the dynamics of social capital, starting with the issue of bridging social capital.

Imagine that all persons are partitioned into a set of groups (say, neighborhoods). We would count ties that go *within* neighborhoods as "bonding" social capital and those that go *between* neighborhoods as "bridging" social capital. The bonding ties have an equivocal nature for us (they might be good for insiders, but bad for outsiders), while the bridging seem an unalloyed good. But how are these bridging ties formed? Perhaps via other associational activities. For example, church co-membership can create bridges connecting those from different neighborhoods (for a related empirical example, see Ruef and Kwon 2016).

But wait a moment! These ties only appear as bridges because we were using the reference frame of neighborhoods, and ignoring all other forms in which persons could be divided up. Had we instead began by considering *religious groups*, these ties between coreligionists in different neighborhoods would appear as *bonding* capital, while those within a neighborhood, but between members of different religious bodies, would appear as the *bridges*. This is a prime example of Simmelian duality—by changing the reference frame (what Simmel would call "turning it on its axis"), our entire evaluative interpretation has turned inside out and upside down, though our core formal structure of group membership data is unchanged. People are not neatly nested in a single set of distinctions: rather, they are simultaneously members of multiple overlapping groups—as Simmel ([1923] 1950) put it best, we are each defined as the intersection of multiple social circles.

It is in part because of this formal instability that there has been so little progress in building any general theory of social capital, and attempts to homogenize all the various uses of the simile of social capital (e.g., Adler and Kwon 2002) could do little more than produce inventories of all the ways that people may have relations, only calling these "capital." But we think that a reconsideration of what might be good about bridging ties, and what might be good about associational memberships, suggests a way to borrow notions from the study of *network* capital—that having interpersonal ties can provide various types of resources for actors.

2.3 What is Good About Groups?

The connection of interest in associational life as a measure of social capital was historically connected to both the Tocquevillian theory of American exceptionalism and the mass society theory of the roots of totalitarianism. It would indeed prove delightful to American history should the two turn into a single theory. But there are already difficulties with the idea that associational social capital brings the claimed results (see Portes and Vickstrom 2011). Indeed, close attention to associational life should provide the last nail in the coffin of what Thomson

(2005) cheekily calls "the theory that wouldn't die," and this is because rather than totalitarianism arising where associational life is weak, as might be derived from classic "mass society" theory (e.g., Kornhauser 1959), both Nazism and Italian fascism grew up in areas in which there were a rich tapestry of associations, precisely because these formed a substrate in which the right-wing movements could spread (e.g., Riley 2010).

That of course does not mean that in other places, perhaps the United States, associational life *isn't* the basis of positive, perhaps even necessary, social capital. Indeed, given the long American love affair with voluntary associations—George Washington's express disapproval of "self-created societies" only put a temporary hold on the explosive growth of American groups (Wood 1992: 329)—it would seem nearly impossible to exaggerate the importance of such group memberships for American social capital. Nevertheless, Putnam (2000) has tried and succeeded in his *Bowling Alone*, down to the inadvertently humorous title, confusing the end of the brief period of *formal associations* around bowling with the beginning of isolation.

Let us use this case to try to figure out what might be so important about associations. If we cannot simply claim that there is some societal-level attribute of having intermediary organizations—that is, that the presence of groups is a *global* measure of the degree of social capital in some place and time—perhaps we may still find that group memberships express *individual* variations in at least one portion of any individual's stock of social capital. The more group memberships any person has, then, all other things being equal, the more we believe them to possess social capital.

We do not deny that there may be some aspects of membership that are themselves important for social capital—members may receive information (e.g., newsletters), access (e.g., museum admission), legitimacy (e.g., professional organizations), and so on. But, as Hooghe and Quintelier (2013) remind us, not all group memberships are the same. Hence the interest in looking not at the total *number* of memberships, but particular forms (e.g., neighborhood groups), the spread across different *types* (e.g., Cigler and Joslyn 2002; Li et al. 2005), or even relationships *between* groups (Oh, Labianca and Chung 2006) on the reasonable assumption that these indicate a range of experience.

Further, we might expect that, in addition to any such benefits, formal associations may also offer the members the chance to establish *explicit* ties to those with whom, as co-members, they already have *implicit* ties. This then suggests the potential for serious errors in past uses of memberships to make arguments about social capital. Anyone who has actually *been* to a bowling alley, as Boggs (2001) notes, knows that *no one* bowls alone—they bowl with *friends*. Associations might be especially important for the friendless—it gives them someone to bowl with, at the cost of membership dues and meetings. This again is to propose a Simmelian intervention: group memberships may be important because they scaffold the creation of ties among otherwise unlinked co-members (see, most importantly, Small 2010). This way of thinking about group membership may have the advantage of requiring few assumptions about the nature of civil society. But it also, as we go on to show, can solve the problem of the formal instability of bridging and bonding capital.

2.4 Redundancy and Group Memberships

The simplest idea of relational social capital is that it is good to have friends, and the somewhat more sophisticated version (held by Bourdieu) is that it is good to have friends in the right places. But, building on the pivotal work of Granovetter (1973), Burt (1992) introduced a wholly structural amendation: it is good to have ties that are *non-redundant*. Those who are enmeshed in dense, highly closed, networks, may have trustworthy confidants, but they also may be stifled by the strong norms of the community, and they will have a hard time getting information that they do not have already. The friend who is friends with your other friends is unlikely to tell you something you don't know, while the friendship that bridges a "structural hole" can give you a first-mover advantage in grappling with new information. (It is not quite this stark; Burt and Merluzzi [2016] argue that best of all is an alternation between the two sorts of network structures.)

If there is a special advantage to non-redundant ties, this would presumably also characterize those ties that are scaffolded by group co-membership. That would imply that while it may be advantageous to belong to groups, it is better when these groups put one in contact with non-redundant alters. This way of thinking allows us to save the valuable insight underlying the notion of bridging capital, by instead focusing on the *diversity* of co-members. This would, we argue, be precisely the approach that would have appealed to John Mohr.

For John Mohr, social groupings were primarily interesting in that they represent distinct cultural worlds; each coalition is united by a set of cultural norms. But Mohr rejected the essentialist vision of a single grid which divides humanity into cultures, cultures into subcultures, and so on. Both in his practice as an administrator struggling to keep education accessible to historically underrepresented groups (Castro, Fenstermaker, Mohr and Guckenheimer 2009) and as an analyst (Mohr and Lee 2000), Mohr focused on the key fact that different categorical schemes incompletely overlapped. Indeed, following Mohr, and taking the idea of "diversity" seriously, we find a way forwards that is free from the paradoxes of "bridging capital." As we saw above, since what is bridging capital according to one scheme is bonding according to another, it makes little sense to propose a general metric of bridging capital. This is not, we will show, true of measuring the *diversity* in social capital. Thus here, we will develop the notion of "Mohr social capital" as specifically that form of capital that leads to access to diversity. All other things being equal, the more relationships, the more distinctive the relationships, and the more balanced one's attention across relationships, the more (Mohr) social capital one has.

2.5 Mass and Diversity

Thus we can make a distinction between two analytic dimensions of associational social capital coming from group memberships, which we shall term *mass* and *diversity*. *Mass* is the total amount of connectivity that group membership facilitates. In some cases, this might be the most important dimension of social capital for accomplishing certain goals. These are goals in which the mere availability of others—no matter whom they are—is useful: for example, putting out the word to look for a lost dog, selling goods, or finding someone to listen to you vent about a problem. *Diversity*, in contrast, is the amount of heterogeneity captured by those memberships and relationships—whether our number of relationships, great or small, present us with a variety of people and ways of thinking, or mostly more of the same. It asks us to take an ecological perspective on social relationships, where relationships represent access to different cultural worlds. We can imagine that some other social goals—e.g., finding a new job, brainstorming a

solution to a complex problem—are well facilitated by the diversity of one's social capital. Of course, most social goals are best served by mass and diversity in some combination, and creating a metric is largely about striking the right balance between these components.

Note that speaking of the *diversity* of social capital does not contain the paradox of attempting to differentiate bonding from bridging capital. And while bridging capital was understood predominantly as a *collective* good, here we build on Burt (1992) to recognize that—as John Mohr believed—diversity can be good for ego as well, as ego becomes exposed to diverse influences.

We propose to distinguish between this *formal* issue of social capital as mass and as diversity for both the substantive realms of Tocquevillian (group-oriented) and Colemanian (individual) social capital (Table 1 places different measures of social capital, indicated by c^1, c^2, and so on, in a two-by-two table following this conceptualization), where the Colemanian capital is specifically that coming from *co-memberships*.[2] To do this, we would need not simply a sample of persons asked about their groups, but the membership rosters of all these groups. Data on such complete membership rosters, however, has, so far as we know, never been used to estimate social capital, until now. We go on to describe the data that we use to do precisely this.

Table 1: Social Capital Measures for Different Conceptions/Forms

		Substantive Characteristics of Social Capital	
		Tocquevillian	*Colemanian*
Formal Characteristics	*Mass*	c^1	c^2, c^3
	Diversity	c^4	c^5, c^6

3. Data

We are interested in studying social capital beginning from the classic Breiger (1974) persons-by-groups matrix. But as McPherson (1982) emphasizes in a wonderful article building on Breiger's approach to duality, the distribution of group sizes is highly skewed (mean above median), with most groups very small and some groups very large. For this reason, how we conceive of our question can greatly affect our results. If we sample on groups treating each group as a unit at risk, we tend to get many small groups, of which very few (and presumably unrepresentative) individuals are members. However, if we sample on individuals, we lose the diversity of groups, as most people are only in a few very large groups. What would be best, of course, is having no need to sample at all.

This is our approach. Our main data consist of information on active U.S. participants in Facebook Groups as of June 30, 2020. Facebook Groups are excellent data with which to analyze the ways that group affiliations connect individuals usefully because they are all

[2]. Here we do not mean to indicate that this social capital involves *closure* as opposed to *openness*, as in Reagans and Zuckerman's (2001) contrast of Colemanian and Burtian social capital; because we begin with two-mode data, the group co-memberships are inherently saturated.

organized at a basic level, yet they still span a continuum of formal and informal organizations. Some groups represent true formal organizations with membership dues and scheduled social commitments (e.g., "Junior Elite Bowling League"), while others emerge from common informal identities (e.g., "Doc's Gang") or common interests (e.g., "Bowling Talk").

There are three great advantages of this data. The first is that these sorts of groups fit the kind of theoretical world we have sketched. They both can facilitate interaction between members (as can face-to-face community groups, but not all formal organizations) while allowing for very diverse co-membership (as do large formal organizations, but not all face-to-face relations). Second, complete membership rosters for all groups are stored on Facebook servers. Third, we have no intrinsic need to sample at all.

That said, we do make a few decisions to ensure that the data we use are maximally internally comparable. First, we only treat as focal egos those who are members of at least two groups (necessary for our diversity scores). Second, we here consider only "active" membership relations, defined by an individual having viewed a group's content at least once in the past seven days. We also only consider groups where at least 50% of the members are located in the United States. This helps us capture behaviors within a single cultural context where Facebook groups are understood as discussion forums uniting people around common interests and identities. The average number of groups to which our users belong is 8.5, and the maximum 552. We thus sample on individuals, but have a complete sample of active users within our constraints. We use anonymized data, preserving only the number of group memberships, frequency of interactions with each group over the past week, gender, age, and county of residence for each person.[3] Our resulting dataset then contains 77,414,956 U.S. user accounts who are, in total, active members of 8,766,915 Facebook groups.

Finally, we are also able to assign all persons to "types" based on the "Social Hash" algorithm (Shalita et al. 2016). Much of the information on Facebook actions, whether we are speaking of the existence of friendship relations, commenting on posts, liking other's posts, and tags, are relational, and stored in a vast network (the "Social Graph"). The Social Hash algorithm is used to partition this graph to increase the efficiency of relational queries and lookups. Given the vast size of the Social Graph, different parts must be stored multiply on different computers, and the speed of making a walk from one part of the graph to another (for example, sending a message to the friend of a friend) is increased if this walk stays within the same unit. The Social Hash algorithm determines the best way of sorting the nodes of the network into a hierarchically nested set of cuts producing buckets of accounts at any cut-level such that edges (Facebook friendship ties) are most likely to be within as opposed to between buckets. At the lowest, buckets may empirically tend to correspond to clusters of friends, co-workers, or members of organizations such as churches or schools. At the higher level cuts we will be using, such buckets are aggregated into larger entities that may tend to be similar in terms of predominant language, age, national/ethnic origin and especially location. This, then, fits our interest in determining, for any ego, the "sorts of people" that ego is at high risk of knowing.

From this model, we do not know all of the *reasons* why any two people do or do not end up in

[3]. Because we sample on individuals, we do not preserve statistics on groups and hence do not give statistics on average group sizes.

the same bucket. Certainly, geographic location is a big part of the story, but anything that leads people to tend to form ties is part of the explanation here. Since our use of this measure is to find those whose group-induced co-membership ties include persons who are unlikely to already be in contact, this lack of clarity as to the reasons for the placement of persons in buckets is nonproblematic. The Social Hash algorithm constructs a hierarchical set of bifurcations minimizing cross-bucket linkages; we here take the buckets that result from the 10^{th} cut, leading to 2^{10}=1,024 total possible types. Any group then has a probability vector across these different types.

4. Mass-Based Measures

We go on to derive measures of these four species of social capital, and illustrate them in two ways. First, we will be using the Facebook data just introduced. Second, we will also at times use an example set of data. We will assume that our data fall in the form of the classic Breiger (1974) person × group data matrix. Imagine that we have N individuals, each of whom can be a member of any or all of M different groups. The data $\mathbf{X}$ is defined $x_{ik} = 1$ if person i is a member of group k and 0 otherwise. Our "toy" example will be $\mathbf{X}$ given in Table 2, containing information on the membership of 6 persons (*A*-*F*) in six groups (*1*-*6*). We summarize our basic arguments about what each measure is and for what sorts of questions it would be appropriate in Table 3.

Table 2: Toy Example

	1	*2*	*3*	*4*	*5*	*6*
A	1	1	1	0	0	0
B	0	0	0	1	0	1
C	0	1	0	0	1	0
D	0	0	0	1	1	0
E	0	0	0	1	0	1
F	0	0	0	0	0	1

Table 3: Summary of Measures

Measure	Substance	Form	Operationalization	Data Needed	Benefits
c^1	Tocquevillian	Mass	Sum of group memberships	Individual survey	General: Broadcast information exposure Example: Receiving broadcast, access, signaling (e.g., Viswanath and Randolph Steele 2006)
c^2	Colemanian	Mass	Weighted sum of group memberships	Individual survey + Organizational information	General: Obtaining social support or other low-cost goods Example: Volunteering, fundraising (e.g., Velthuis 2017)
c^3	Colemanian	Mass	Non-redundant co-members	Complete organizational rosters	Same as c^2; superior measure (e.g., Burt, 1992)
c^4	Tocquevillian	Diversity	Rao-Stirling; Endogenous measures	Complete organizational rosters	General: Access a variety of perspectives/opinions Example: Gaining perspective, elaboration of personality, brainstorming a solution to an interdisciplinary problem (e.g., Leydesdorff and Rafols 2011)
c^5	Colemanian	Diversity	Group co-member entropy	Complete organizational rosters	General: Knowledge gain, particularly where knowledge is obscure Example: Information (e.g., Li et al. 2005) Special case: Benefits related to social cohesion by comparing within and between diversities (e.g., Fieldhouse and Cutts 2010)
c^6	Colemanian	Diversity	Between-groups entropy	Complete organizational rosters	General: Synthesize knowledge from multiple perspectives Example: Population inference (e.g., Kurzman, 2004), political tolerance (e.g., Gigler and Joslyn 2002)

Note: Here, we focus on information/knowledge/exposure benefit. We ignore cultural and cognitive aspects and also tie strengths that can address various types of benefits and network/participation cost, jointly with our conception.

4.1 Tocquevillian Mass (*Measure 1*)

First, let us consider mass-based measures of Tocquevillian social capital. The most obvious such measure (c^1) is simply the number of groups to which anyone belongs:

$$c^1_i = \sum_k x_{ik} \quad (1)$$

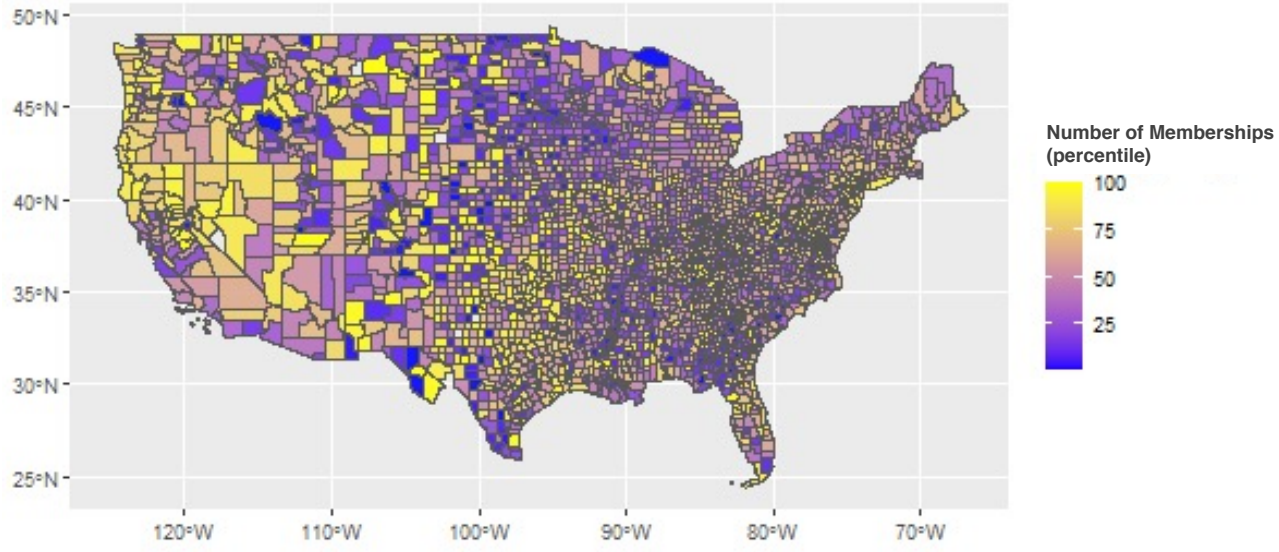
In other words, we are taking the column sums of our data matrix. For our toy example (Table 2), by c^1, *A* has the most social capital, as she is a member of 3 groups, while everyone else has only 1 or 2 memberships.

A consideration of Tocquevillian mass may be most appropriate when we are interested in the *access to broadcast information* that persons might have via formal group memberships. Each membership represents access to the information broadcast in that group.

An example derived from our Facebook data is mapped below (Figure 1), showing the average number of groups to which our American members belong, organized by county. (Here and in the following graphs, these statistics are turned into percentiles to facilitate visualization.) This is a relatively easy measure to construct even without Facebook data, as we can in most circumstances ascertain each individual's total number of groups simply by asking in a survey.

We see a band of high social capital that might not be where we first imagine it, one running down Appalachia, and another in the lower plains. In particular, the western Mountain region appears to be the place of great social capital.

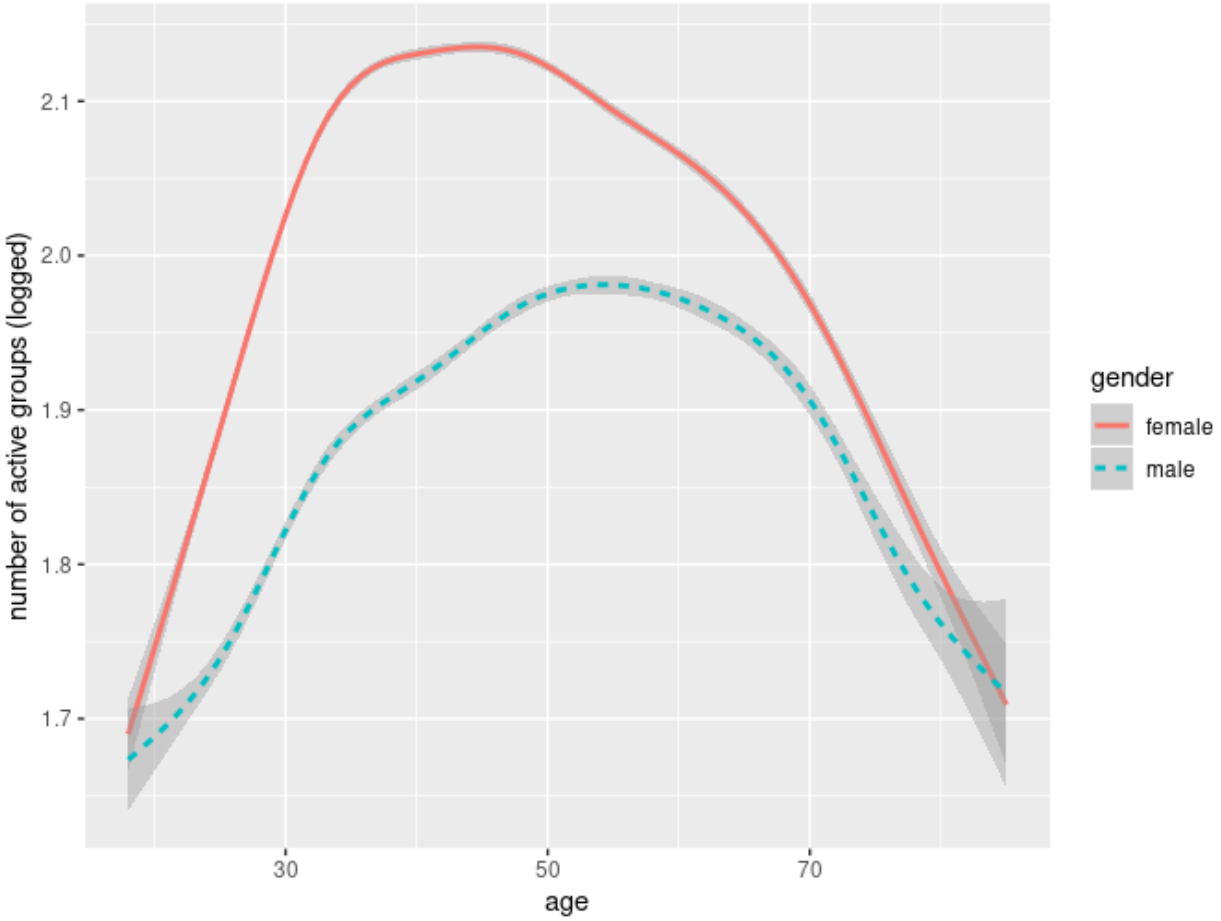
Figure 1: Active Group Memberships by County



When we look at this measure of social capital by gender and age (limiting age to those between 18 and 85; number of groups logged), we see that social capital is greater for women than men consistently across all age groups—women consistently join more Facebook groups than men. We also see that the age of maximum social capital is lower for women than for men. Women in their mid 40s and men in their mid 50s have the most group memberships in their respective

gender groups.

Figure 2: Active Memberships by Age and Gender



But even when we consider only the issue of information, group membership can be important not only because of information broadcasts, but also because of the number of individuals that a focal member is put in touch with as co-members. Thus we turn here to measures of Colemanian mass.

4.2 Colemanian Mass: Redundant (*Measure 2*)

The number of groups in which one is a member makes intuitive sense as a measure of social capital because it represents the raw number of arenas in which one can interact with others. But this measure does not differentiate between membership in small vs. large groups. Persons A and B may both be members of 8 groups, but person A's groups have 2 members each and person B's groups have 300 members each. It is reasonable in this case to propose that person B actually has more social capital.

This may initially seem counter-intuitive—we tend to imagine that when it comes to the sorts of

relationships and actions that strike us as most paradigmatically social capital-ish, smaller, *gemeinschaftliche* groups will be far stronger than large, anonymous, *gesellschaftliche* ones. Without denying that there can be a great deal of truth in this, it is easy to underestimate the magnitude of magnitude, as it were. If one is trying to get a kidney transplant from another group member, since one only needs one donor (assuming compatibility), if all group members in a group of 10 had a 1-in-10 chance of volunteering, ego's chance of getting a match is around 2/3. But if one is soliciting from a group of 1000, one has around the same probability of a match if the others have a 1-in-1000 chance of volunteering.[4] The same may be true for certain forms of information. For example, when a person considers relocating to a new city, the person may join a number of local Facebook groups in order to collect a wide range of both information and also *opinions* regarding neighborhoods and their differences in terms of amenities, or the quality of needed replacements for existing services and relations (for example, childcare, recreation, arts). Those extensive memberships may be temporal, ephemeral, and contingent, but still no reason to assume that those memberships are not meaningful. Most important, ego may find both being able to reach multiple people who are willing to give either different points of views (can children ride bikes safely?) and the ability to address relatively obscure interests (is there a good place for Bocce ball?) to be more successful in large groups than in small.

Access to many co-members can also be advantageous where ego is attempting to use contacts for fundraising—whether asking for voluntary contributions (say, to a health fund) or selling goods. Of course, we might imagine that smaller groups will have an advantage where the *negative* sanction of group disapproval is a motivator (which entails that donations be visible to all others). But there are also cases in which actors on social media seem motivated to make contributions for the *positive* esteem that they receive from others, even where they do not have relations with those who witness the contribution (e.g., Velthuis 2017). Thus we suggest that *for social goods that either are low-cost, or, if costly, need few providers, and where positive sanctions are more important than negative ones in securing participation, large groups may indeed provide more social capital than small groups.*

Moving to try to measure the degree to which ego is put in contact with others, then, we want to take into account the mass of group size. Let us denote the sum of the i^{th} row of any matrix $\mathbf{X}$ $x_{i\bullet} = \sum_k x_{ik}$ and the sum of the k^{th} column $\mathbf{X}$ $x_{\bullet k} = \sum x_{ik}$. We might, therefore, consider ego's social capital to be a function of the group sizes of the groups of which she is a member. Thus we can propose

$$c^2_i = \sum_k x_{ik} \left(x_{\bullet k} - 1 \right) \tag{2}$$

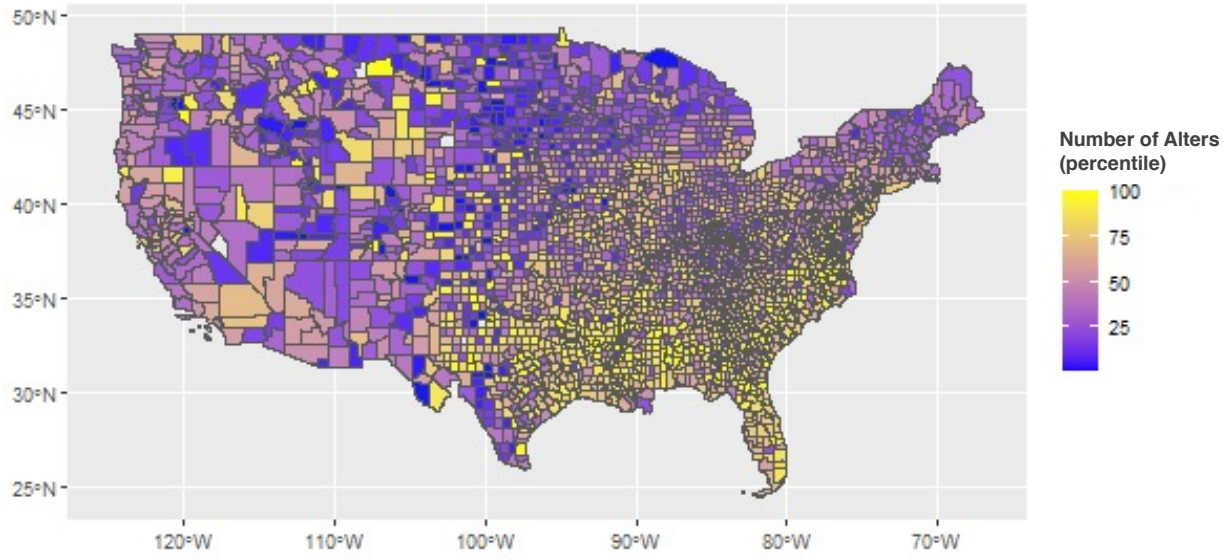
as the sum of the group sizes of the groups to which *i* belongs (subtracting 1 for ego). For the data in Table 2, persons *B* and *E* would have the highest social capital, as they are members of two groups of size 3; hence $c^2 = (3 - 1) + (3 - 1) = 4$. We will note below why this is actually not an exact operationalization of our theoretical interest in Colemanian mass, but we include it because it can be produced using only a combination of a survey of individuals (what groups do you belong to) and organizational information on group sizes.

---

[4]. While in this case, the relations seem linear, they aren't really: given a probability of any one person volunteering of p, and a group size of N, the chance that ego gets a match $P = 1 - (1 - p)^N$.

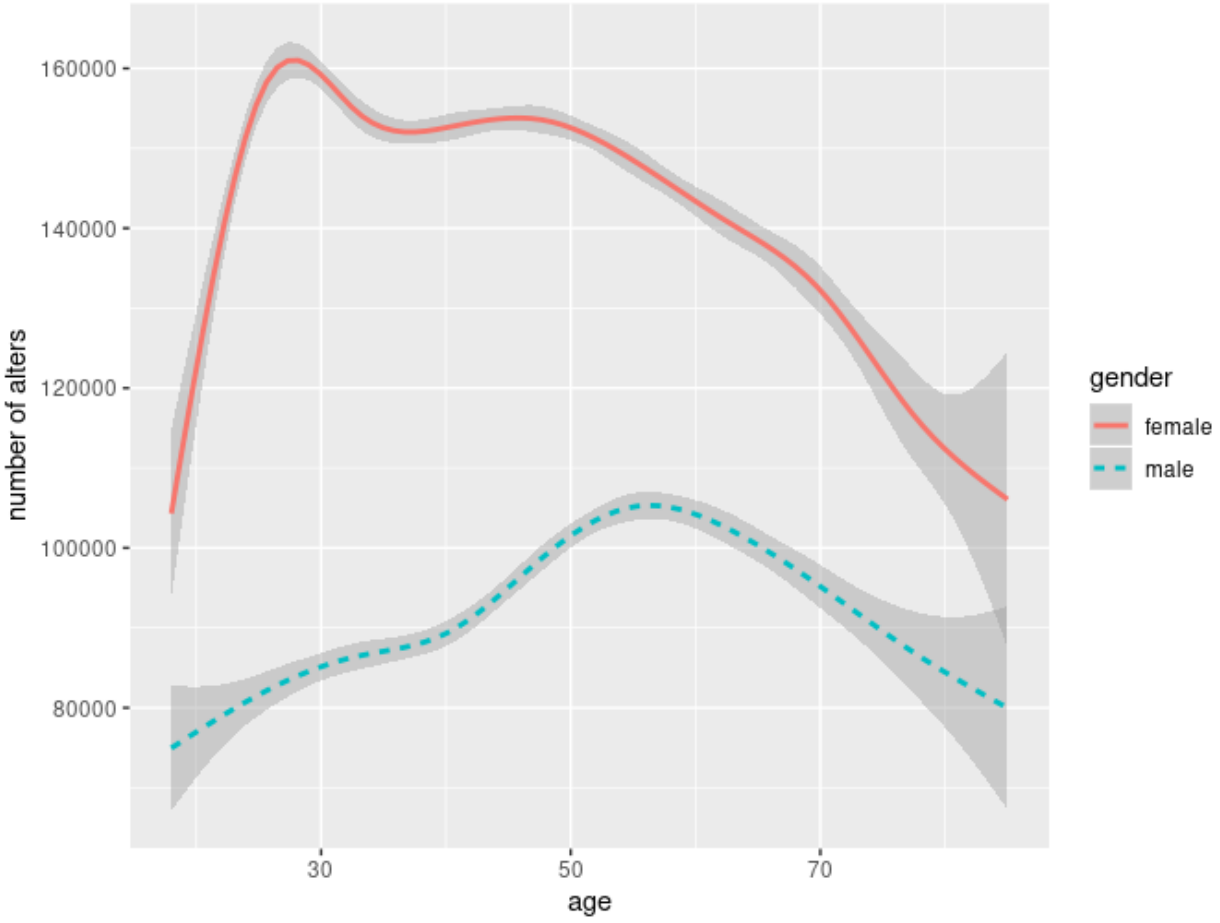
This appears a radically different measure of social capital, despite, like the previous, being mass-based. For when we examine the geographical and demographic distribution of this measure. In particular, the Mountain states no longer have much social capital, and it is instead concentrated in the Southern Atlantic!

Figure 3: Weighted Memberships by County



We find a striking change also in the gender/age patterns. While women still have higher social capital than do men, the shape of the curve is quite different, with women in their late 20s having vastly higher amounts than women in their mid 40s. This means that women in their mid twenties tend to be members of *fewer, larger* groups, whereas middle aged women tend to be in *more, smaller* groups. There is no analogous difference for men. Curiously, the curve for men remains largely the same, where we see an inverted U-shape curve that peaks for men in their mid 50s.

Figure 4: Weighted Memberships by Age and Gender



4.3 Colemanian Mass: Non-Redundant (*Measure 3*)

Of course, with the aggregate data that must be used in such cases, we do not know whether two groups that are formally separate may connect the same individuals. Those who have studied the ways in which social movement organizations are linked by co-memberships (most importantly, Mario Diani [e.g., 2003] and Ann Mische [2008]) may find that the diversity of group memberships can be misleading because the same twenty people may be at the core of fifteen different groups with apparently wildly different foci ("Animal Rights Now!" and "Leftist Socialist Workers Party"). Thus actually understanding the degree of social ties facilitated by any membership requires that we know not merely what groups someone belongs to, but who *else* belongs to such groups.

While in some cases—for example, when we are thinking about costly behaviors, as opposed to information-dissemination—it might be that the redundancy of co-memberships *increases* social capital, as these are akin to "multiplex ties," we think there are, especially for the case of Facebook groups and for formal organizations, relatively few reasons to think that such multiplexity is significant. We therefore think a better measure of Colemanian ties is one that goes in the opposite direction, and makes sure not to count ego as having seven additional ties if

ego is in seven groups with the same alter.

This brings us to a superior measure of social capital in terms of the Colemanian mass: the number of *non-redundant* social relationships facilitated by group membership. Define the "transpose" of $\mathbf{X}$ to be a matrix $\mathbf{X}^t$ such that the $(i,k)^{th}$ element of $\mathbf{X}^t$ (denoted x^t_{ik}) is the same as the $(k,i)^{th}$ element of $\mathbf{X}$ (x_{ki}). In other words, we flip the matrix on its diagonal. Following Simmel, Breiger (1974) suggested that we use simple matrix multiplication to derive the pattern of shared group memberships between persons, and the shared number of co-members between groups. Regarding the first, the $(i,j)^{th}$ element of the matrix $\mathbf{X^*} = \mathbf{XX}^t$ contains the number of groups that person i and person j both are members of; regarding the second, the $(k,h)^{th}$ element of the matrix $\mathbf{X^{**}} = \mathbf{X}^t\mathbf{X}$ contains the number of persons who are both in group k and in group h. If we define our operations as Boolean (1 + 1 = 1), then these matrices contain not the *number* of co-memberships, but rather the logical answer to the question of whether there are *any* co-holdings. Thus to use the example data from Table 2 above, we would construct:

$\mathbf{X^*}$ =

	A	*B*	*C*	*D*	*E*	*F*
A	0	0	1	0	0	0
B	0	0	0	1	1	1
C	1	0	0	1	0	0
D	0	1	1	0	1	0
E	0	1	0	1	0	1
F	0	1	0	0	1	0

and $\mathbf{X^{**}}$ =

	1	*2*	*3*	*4*	*5*	*6*
1	0	1	1	0	0	0
2	1	0	1	0	1	1
3	1	1	0	0	1	0
4	0	0	0	0	1	1
5	0	1	0	1	0	0
6	0	0	0	1	0	0

Also note that this constructed matrix is a close analogue to a correlation matrix: the $(k,h)^{th}$ element of the matrix $\mathbf{X^{**}} = \mathbf{X}^t\mathbf{X}$ is equal to the dot product $\sum_i x_{ik}x_{ih}$, in turn equal to the

correlation of vectors if these vectors are standardized to have unit magnitude. Thus, we can propose a measure of group-relevant social capital that counts the number of unique persons to whom *i* is tied by co-membership:

$$c^3_i = \sum_{k,k\neq i} \mathbf{X}^*_{ik} = \sum_{k,k\neq i} \left(\mathbf{X}\mathbf{X}^t\right)_{ik} \quad (3)$$

Turning to the Facebook data, both the geographical and age/gender distributions look extremely close to those that we portrayed in Figures 3 and 4 (for this reason, we do not reproduce them here). This suggests that, despite the fact that many of our groups are small, the capacity of ego to be in more than one group with alter does not shape the overall distribution of social capital.

4.4 Conclusion to Mass-Based Measures

We believe that for some sorts of questions, mass-based measures of social capital may be most important—the more Facebook groups on vaccines one belongs to, perhaps the more likely one is to hear a vital piece of information ("they post new appointments on the ':02's!"); the more co-members one can reach, the more likely one is to find *someone* who has preserved videotapes of the original Uncle Floyd Show. But for other issues, perhaps involving information, paradigmatically attempts to assess the social whole (e.g., are most Americans doing better or worse?), it can be just as important that one have *diversity* in one's capital, whether Tocquevillian or Colemanian. We go on to introduce some new measures of the diversity of social capital.

5. Diversity

The core insight in computing diversity-based measures extends the notion of non-redundancy that we first raised when considering how best to measure the mass of Colemanian social capital arising from group memberships. There, we wanted to avoid repeatedly counting the same alter who may be co-members with ego in multiple groups. But now we want to generalize to think about *kinds* of groups and *kinds* of alters. The guiding intuition is that being a member of two groups, or being a co-member with two alters, contributes less to one's social capital when those two groups or persons are similar in some way to be determined. Those with more diverse ties may be able to draw upon more sources of information and perspective than those with more homogenous ties (e.g., Reagans and Zuckerman 2001).

There are two basic ways that we can attempt to ascertain the similarity of two groups or members, one *endogenous* to the basic data table **X** and the other *exogenous*. The former follows the extended Mohr/Breiger approach to social data taken by Kovács (2010), Lizardo (2018) and Lee and Martin (2018) in recycling information about the similarity of rows when creating measures for columns, and information about the similarity of columns when creating measures for rows. The latter uses data produced by Facebook to predict tie formation between persons. While one could use either an endogenous or an exogenous approach for either Tocquevillian or Colemanian social capital, in the interests of parsimony (as well as different operationalizations of diversity), here we will work through measures of endogenously generated Tocquevillian diversity and exogenously generated Colemanian diversity. We begin with the former.

5.1 Tocquevillian Diversity (*Measure 4*)

If what groups offer their members *qua* groups (and not via implicit ties to co-members) is broadcast information and authorized access, why does it matter whether the groups are similar to one another? For one, a greater variety of group types (perhaps the sort of thing best measured exogenously) can offer ego multiple points of entry to the social whole—more possibilities of exploration and self-development. But we also suggest that one important form of social capital that groups offer members comes in the form of the provision of a virtual "place to stand" from which to take a perspective. While in some cases this might have direct implications for how one approaches a problem ("as a member of the Oregon Society of American Foresters, I see this issue one way, but as a member of the Sierra Club, I see it somewhat differently"), we also think that this might be, more subtly, a way that persons cultivate and energize different facets of their personality.

And in this more subtle sense, we propose that the diversity, as opposed to the overlap, of members, may be very important for whether different group memberships really do provide ego with a full panoply of these imagined "places to stand." The core notion of the Breiger-Mohr duality approach to groups is based on the formalization of people as intersections of their groups, and groups as unions of their members. Groups that have the same members are expected to be, informally if not formally, in some way similar, and when all our groups share members, we are not "individuals" in the Durkheimian sense. Hence the interest in a specifically endogenous measure of the diversity of Tocquevillian social capital.

Let us begin by attempting to create a *dissimilarity* score d_{gh} for any two groups g and h, and accordingly a matrix **D** of all dissimilarity scores between groups. The only constraint we place on this measure is for convenience that $0 \leq d_{gh} \leq 1$. We could of course use exogenous information to compute this dissimilarity (e.g., if groups had been coded on theme, organizational structure, nonprofit/for-profit, and so on). But, following the logic laid out above, here we illustrate the use of an endogenous approach to the computation of Tocquevillian diversity. We can use our group membership matrix to calculate the similarity or dissimilarity between groups by assuming that groups that tend to share members are more similar than groups whose memberships overlap less.

Let the dissimilarity d_{kh} between group k and group h, from the perspective of group k, be the proportion of k's members who are *not* also in h; similarly, let the dissimilarity d_{hk} between group k and group h, from the perspective of group h, be the proportion of h's members who are *not* also in k. Formally, define $n_{kh} = \sum_i x_{ik}x_{ih}$, the number of group members in both group k and group h; for ease of notation now let $n_k = x_{\bullet k}$ and similarly for h. Then $d_{kh} = (1 - n_{kh}/n_k)$ and $d_{hk} = (1 - n_{kh}/n_h)$. This is equivalent to a Jaccard score, but one that is asymmetric.

Note, therefore, that in many cases, $d_{kh} \neq d_{hk}$. For the toy data in Table 2, our **D** matrix looks like this:

	1	*2*	*3*	*4*	*5*	*6*
1	0	0	0	1	1	1
2	½	0	½	1	½	1
3	0	0	0	1	1	1
4	1	1	1	0	2/3	1/3
5	1	½	1	½	0	1
6	1	1	1	1/3	1	0

The (2,1) cell is ½ because ½ of group *2*'s members are not also in group *1* (this would be person *C*), and the (1,2) cell is 0 because none of group *1*'s members are not also in *2*—there is no dissimilarity.

With this measure of group differences, we propose to measure diversity using the widely known Rao-Stirling measure (Stirling 1998; Park et al. 2015) which is often understood as combining variety, balance and disparity of investments across groups. This is usually defined as follows:

$$c^4_i = \sum_{k,h,\, k \neq h} z_{ik} z_{ih} d_{kh} \tag{4}$$

where z_{ik} is a measure of the proportion of person *i's* attention that is allocated to group *k*. In Appendix 1, we discuss more general ways of understanding the joint attention here parameterized as $z_{ik}z_{ih}$.

Most usages of the Rao-Stirling in the computational sciences construct the matrix **D** to be symmetrical, using cosine similarity (Leydesdorff and Rafols 2011) or a normalized Jaccard similarity (Shi, Lim, and Suh 2018), to express the member-similarity between groups. The Jaccard might be normalized by using the size of the smaller of the two groups, which has, for instance, been shown a useful way to express the 'unconventionality' value of combining two sets of entities of vastly unequal size (Silver, Lee, Childress 2016). However, mathematical biologists using information theory often find it necessary to treat dissimilarities as asymmetric, and to preserve generality, we make no constraints. In this case, allowing the contribution to the overall Rao-Stirling diversity for ego *i* from cell (2,1) to be different from that coming from cell (1,2) is equivalent to what would be computed were we to average two different normalizations, one which we divide n_{kh} by $\max(n_k, n_h)$ and the other in which we divide it by $\min(n_k, n_h)$.

Thus **D**; we now need to consider how to compute each person's vector of engagement across groups, $\mathbf{z}_i$. If we have data y_{ik} that expresses the degree of engagement that person *i* has with group *k*, we can use this to create an $N \times M$ row-normalized matrix **Z** where $z_{ik} = y_{ik} / \sum_j y_{ij}$. (Row-normalization is not necessary but conceptually clearer and allows a continuity with previous work by others.) The engagement data we use here is the proportion of a user's Facebook group content views that fell on the group in question in the past 7 days, expressed as a number between 0 and 1.

We then create a Rao-Stirling diversity measure of social capital by using these engagement scores and the distances based on membership overlaps. Figure 5 shows the results when we use

this endogenous diversity measure of Tocquevillian social capital. Now we find that it is no longer the case that the Southeast has a surplus of social capital. Once we take out the *mass* of social capital, and just look at diversity, social capital seems consolidated in the Western mountain states and along the East and West coasts.

Figure 5: Tocquevillian Diversity by County

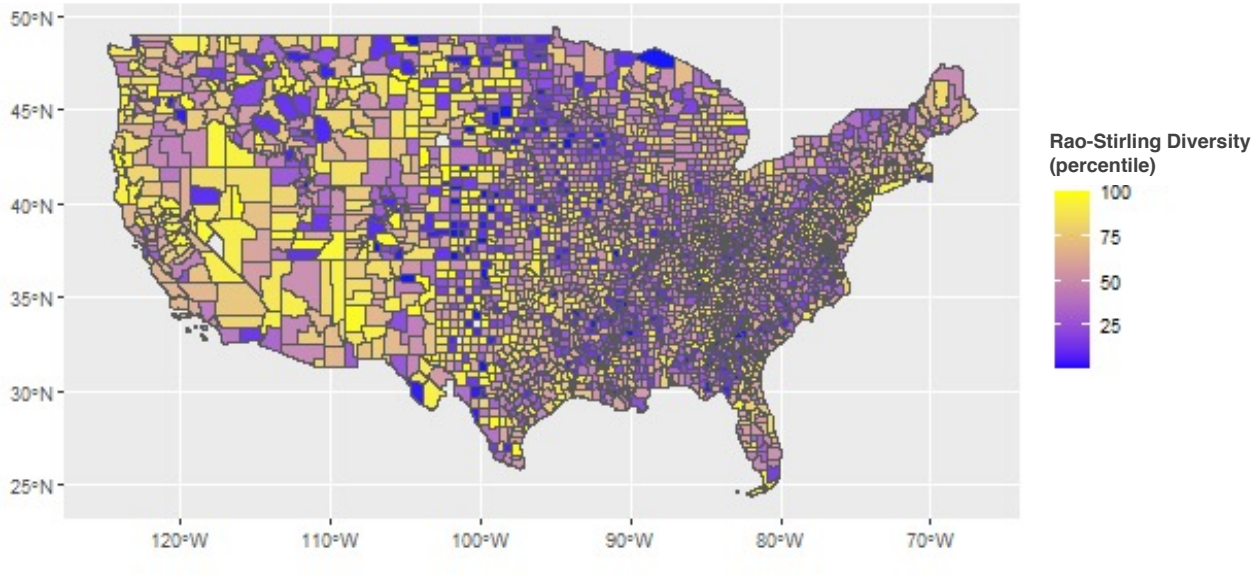
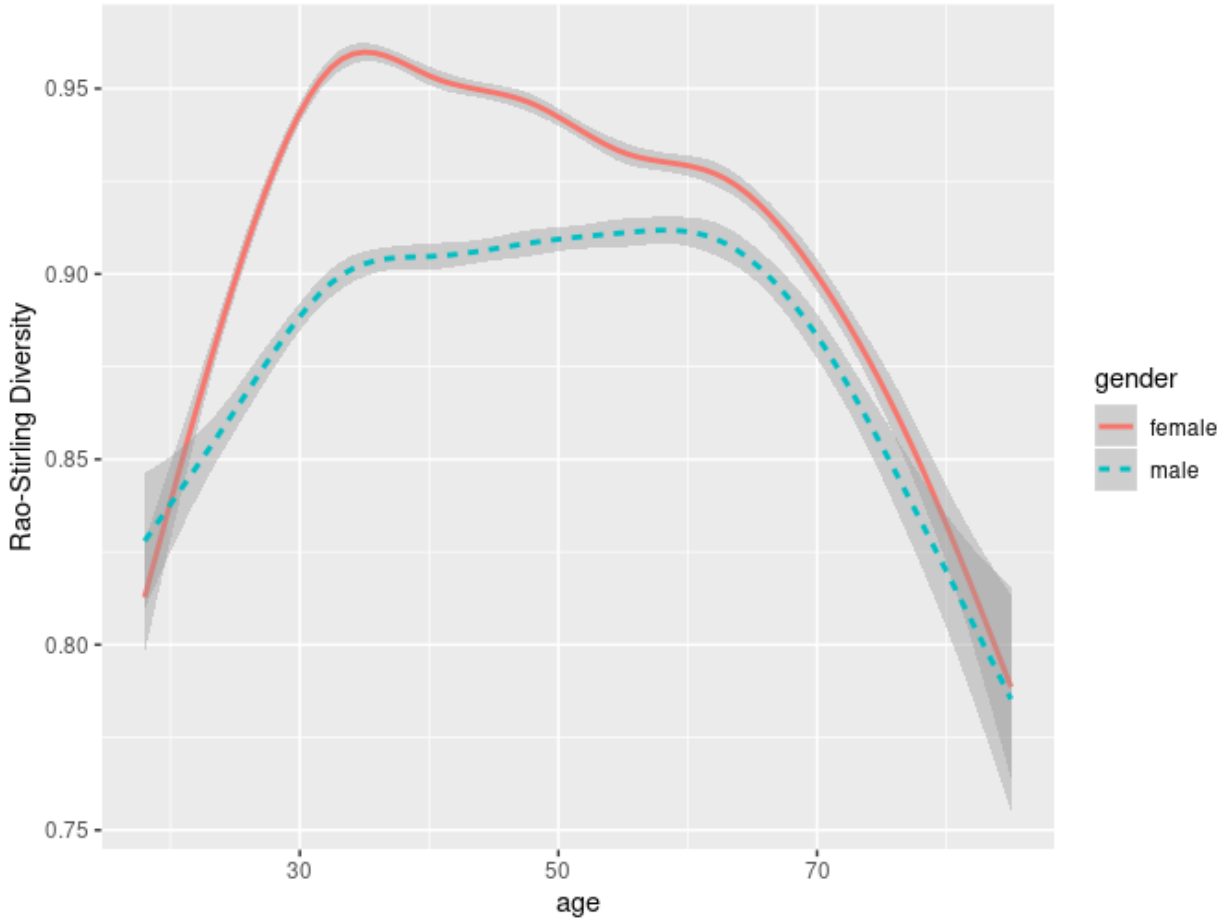


Figure 6: Tocquevillian Diversity by Age and Gender



We also see a change in the gender/age breakdown (see Figure 6). For the first time in our analysis, women in their mid-30s have the highest social capital, in contrast to our previous results which found women in their 20s to have the most co-memberships, and women in their 40s to have the most group memberships. It also may be interesting that our conclusions about men might also change somewhat if we were to consider the diversity, and not the mass, of Tocquevillian social capital. Although the maximum observed has shifted slightly to around age 60, there is a rapid decrease with age from there (which is also seen among women). Thus our conclusions differ by whether we focus on the mass or the diversity of social capital. We go on to show even starker differences.

5.2 Exogenous Measures of Colemanian Dissimilarity (*Measure 5*)

We have used the empirically observed overlaps of the memberships of different groups to create pairwise dissimilarities, which we then use to build a measure of social capital. The guiding notion is that adding a group membership does little for one's social capital unless it increases

the diversity of one's access to the social whole, at least, compared to the groups one is most involved/invested in. The same is likely to be true for the social capital that comes in the form of induced ties between co-members. In particular, we might imagine that the diversity of ties is especially important when it comes to using ties as sources of *information* (for example, by following group discussions). This implies that an additional membership does little to increase one's social capital, even if it establishes implicit ties between unknown alters, if those alters are very much *like* the friends one already has, where likeness is based on some additional data (other than the group memberships). Of course, we could take an endogenous approach to classifying the differences between persons, one that is dual to the approach taken for Tocquevillian diversity, by considering people to be unlike to the extent that they are members of groups that do not share members. But here we illustrate the use of an exogenous measure.

Let us propose that every person in our data has an observed type t ($t \in \{1, 2, \ldots, T\}$). (Here we will be employing the Social Hash algorithm discussed in the Data section.) One group k is then intrinsically more diverse than another k' if it has a broader distribution of persons of different types. Two groups that have no overlap of members may now be considered still to have high similarity if they have similar distribution of *types* of members.

This is formally similar to a problem in ecology: how to quantify the species diversity of different areas. We here follow an approach common therein which uses an entropic measure of diversity. We here give notation that formalizes how we transform our original data into the resulting data for the computation of diversity; we will rely upon this notation again below. What is central is that for each person, we will create a vector **p** that contains the distribution of his or her friends across our T types. Let **A** be a binary $N \times T$ matrix defined $\{a_{it} = 1$ if person i is in type t, 0 otherwise$\}$; then construct $\mathbf{B^*} = (\mathbf{X}^t\mathbf{A})^t = \{b_{tk}\}$, the $T \times M$ matrix containing the number of persons of any type in each group. However, we will be interested in computing the diversity for *each individual* in our sample. For each individual i, construct the $M \times M$ matrix $\mathbf{C}^i$ defined

$$\begin{aligned} c^i_{kk} &= x_{ik,} \\ c^i_{kg} &= 0, k \neq g \end{aligned} \tag{5}$$

In other words, $\mathbf{C}^i$ is a diagonal matrix (all non-diagonal elements are 0) which contains this person i's row in the original person-by-group matrix **X**. With this, we construct $\mathbf{B}^i = \mathbf{B^*} \times \mathbf{C}^i$. In other words, this may be understood as the subset of $\mathbf{B^*}$ containing only the groups of which person i is a member. Because our exposition is always in terms of any particular ego i, to simplify we will drop any notation for the ego in question, and thus here simply speak of **B**, where b_{tk} is the number of individuals of type t of T categories in group k of M groups of which person i is a member, with row sums $b_{t\bullet} = \sum_k b_{tk}$; column sums $b_{\bullet k} = \sum_t b_{tk}$; and total number of observations $V = \sum_t b_{t\bullet} = \sum_k b_{\bullet k}$. Let the overall probability $p_{tk} = b_{tk} / V$; and then the column and row normalized probabilities be (respectively) $p_{t|k=k'} = b_{tk}/b_{\bullet k'}$ and $p_{k|t=t'} = b_{tk} / b_{t'\bullet}$. Then construct the row and column probability vectors respectively $\mathbf{r_{t'}} = \{p_{1|t=t'}, p_{2|t=t'}, \ldots, p_{M|t=t'}\}$, and $\mathbf{c_{k'}} = \{p_{1|k=k'}, p_{2|k=k'}, \ldots, p_{T|k=k'}\}$. Finally, construct the marginal probability vectors $\mathbf{p} \mid p_t = b_{t\bullet}/V$ and $\mathbf{q} \mid q_k = b_{\bullet k}/V$; and note $\sum_t p_t = \sum_k q_k = 1$. For any probability vector such as $\mathbf{p} = \{p_1, p_2, \ldots, p_T\}$, the entropy $H(\mathbf{p})$ is defined:

$$H(\mathbf{p}) = -\sum_t p_t ln(p_t) \tag{6}$$

When attempting to interpret diversity, it is conventional to examine not $H(\mathbf{p})$ but $\exp[H(\mathbf{p})]$ (Jost 2007). This has interpretability in a metric of the number of effective types; on the other hand, the entropy is familiar and easy to manipulate so we will use this metric when appropriate. By convention, $p_t \ln(p_t) = 0$ when $p_t = 0$; thus when the distribution is maximally concentrated (all observations in one state), $H(\mathbf{p}) = 0$; when the distribution maximally dispersed ($p_t = 1/T \; \forall \; t$), $H(\mathbf{p}) = -T(1/T)\ln(1/T) = \ln(T)$. Note that given the definition of $\mathbf{p}$ above this is the total diversity when we do not consider the fact that our sample is broken down into groups.

Thus we derive our fifth measure of social capital:

$$c^5{}_i = \exp\left[H(\mathbf{p}_i)\right] \tag{7}$$

Figure 7 demonstrates that when we replicate our analyses now regarding the distribution of social capital across age/gender using this measure, we come to very different conclusions. Using the Rao-Stirling measure (c^4), previously, we had found that women around 30 seemed to have the greater degree of Mohr social capital. Instead, here (Figure 7) we find that it is *men* who have the greatest diversity. While women appear to have greater endogenous Tocquevillian diversity than do men—they tend to spread their energy across groups that do not share members—these groups do not necessarily have as members different "types" of people as measured by the Social Hash algorithm. This is not because there is a simple relationship between entropy and the number of groups of which one is a member, or the total number of alters (the correlations between the entropy measures and these measures are all $.2 < r < .3$, and the patterns are somewhat curvilinear).

Figure 7: Colemanian Diversity by Age and Gender

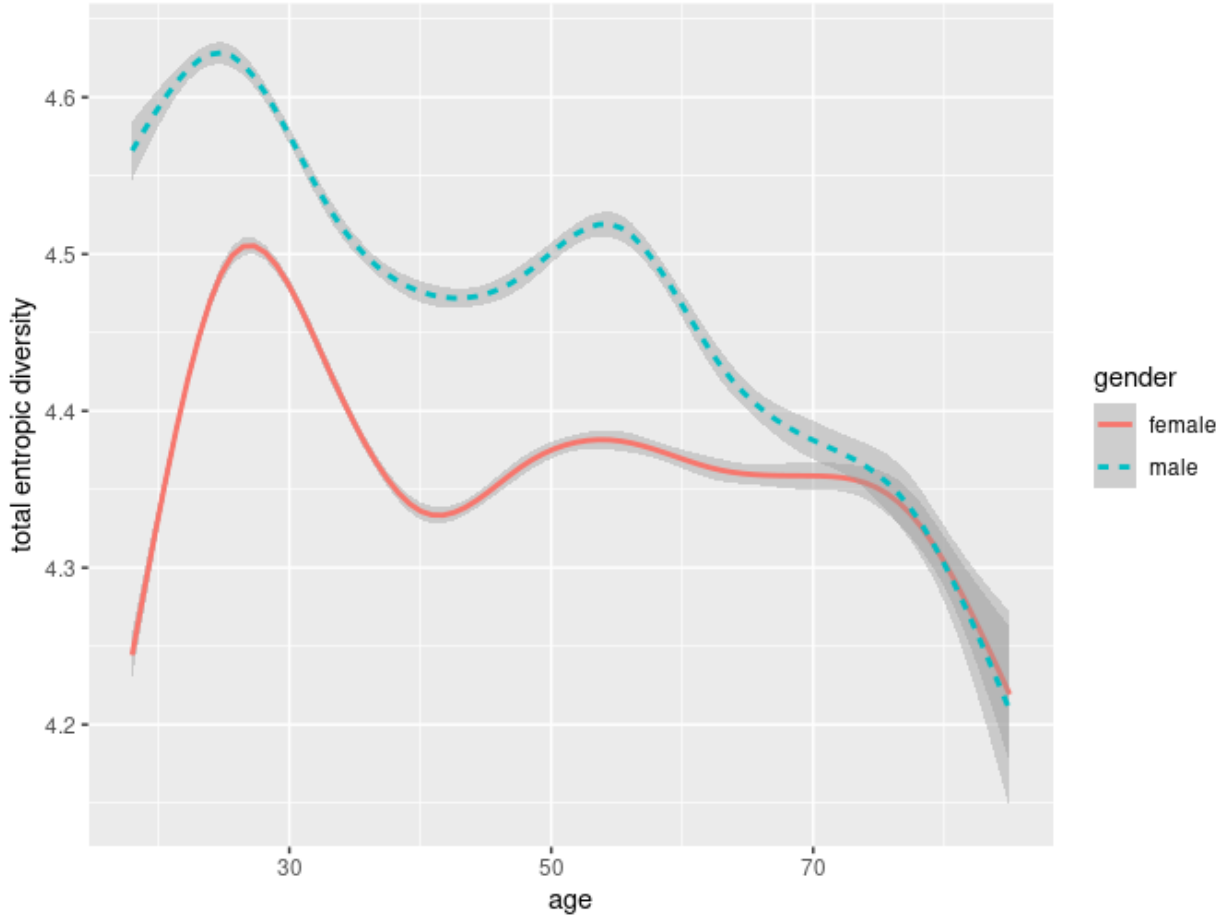
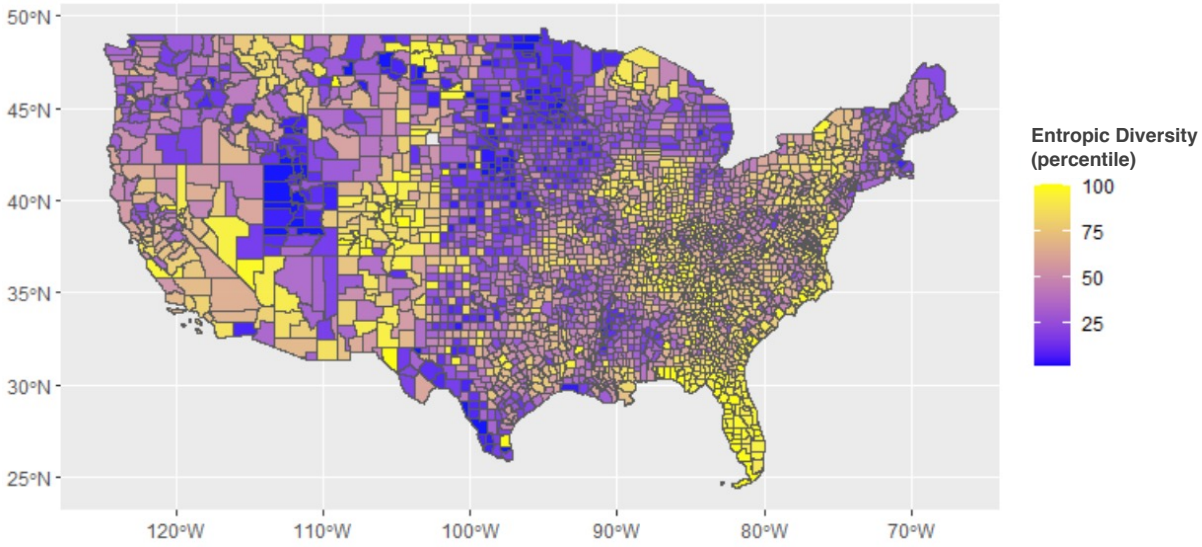


Figure 8: Colemanian Diversity by County



When we look at the geographical spread of this diversity measure (Figure 8), we find that the results are indeed also different from what we had previously seen: this sort of diversity seems concentrated in vacation/retirement states, which is surprising given that we have seen that it tends to decrease with age. We think that this "snowbird" effect is likely to arise because geographical movement is an excellent way of carrying out (seemingly) "random" rewiring in a social network (Martin 2009). People who relocate, especially later in life, may have a bundle of ties to those in different locations that they maintain in their new place, along with new ties to locals, leading to a diverse set of ties. It is also interesting that Utah—along with most of the "Boshington" corridor—has quite low social capital in this sense, despite generally high rates of group involvement.

5.3 Between-Group Colemanian Diversity (*Measure 6*)

We have computed a diversity measure that, in effect, takes all the co-members out of the different groups that they are in, and shuffles them all together. We therefore cannot tell the difference between a person who has diversity because she belongs to groups that contain a great deal of diversity, or because she has a diverse portfolio of internally non-diverse groups. This difference might be extremely consequential: someone who has the "diverse portfolio" of groups may be able to get not simply useful information (for example, codes that can be used to jump ahead in the vaccine line) but a combination of breadth and depth in understanding the variety of perspectives and ideas floating around. When people want to make significant inferences (e.g., in Kurzman's (2004) case, whether dissatisfaction with the Shah is widespread enough to make supporting a rebellion a non-suicidal act), it is important that they be able to leave their social networks and find not just persons of different types, but persons having the conversations that are relevant to different communities. For this reason, we would expect that someone with the diverse portfolio of non-diverse groups may be better suited to make population inferences than the one who belongs to many similarly diverse groups.

An ego can have a diverse set of co-members in two archetypically different ways. First, ego could be the member of groups that are themselves diverse. Indeed, ego could have maximum diversity by belonging to a single group, so long as that group itself was maximally diverse. For example, if $T = 3$, and we write the proportion of members of group k who are type t as $p_{t|k}$, then if ego is a member of a single group $k = 1$, and $p_{\mathbf{t}|k=1} = \{1/3, 1/3, 1/3\}$, ego has maximal diversity. Or ego could have a portfolio of memberships that are in groups that internally are quite homogenous, but different from one another; this an ego who belonged to three equally sized groups with type-breakdowns of {1, 0, 0}, {0, 1, 0}, and {0, 0, 1} would have the same overall diversity as the previous ego.

Any total diversity can therefore be partitioned into two portions, one solely *within* groups and the other *between* groups. Our first ego's diversity is wholly within groups, the second's, wholly between. We go on to consider how to quantify and decompose this approach. Fortunately, this problem is formally homologous to one considered in ecology, where it has proven important to quantify the species-diversity of different areas (akin to our persons) on the basis of different samples from different subareas (akin to our groups). Because some of these results are relatively new, and because many social scientists are not familiar with this literature, we work through certain fundamental results as an appendix here.

The relevant problem is how to quantify the species diversity of an ecosystem from which one has taken several samples (say, from qualitatively different areas). Whittaker's (1972) notion, which has guided subsequent thinking, is that the total ("gamma") diversity (D_γ) should be decomposable into two portions, one having to do with the diversity *within* the samples ("alpha"; D_α), and the other, the diversity *among* (*between*) the samples ("beta"; D_β). We go on to use this approach to determine the diversity of users' profiles of group memberships.

We therefore want to separate the diversity that comes in these two forms. The entropy is one of a number of ways of quantifying diversity, but it has a unique characteristic of being decomposable to within- and between-group components that is lacked by some other diversity scores, even though all of these can be considered members of a larger family; we refer the interested reader to Jost (2006).

We begin by constructing the within-group diversity. The entropy already gives us the diversity of any particular group; we only need to weight the column entropies by their contribution to the overall sample such that the weights sum to 1. Although these weights may be chosen in any way, here we assume that they are proportional to the group sizes, and thus are = **q**. Accordingly, D_α is then

$$D_\alpha = -\sum_k q_k \sum_t p_{t|k} ln\left(p_{t|k}\right) \tag{8}$$

$$= -\sum_k \frac{b_{\bullet k}}{V} \sum_t \frac{b_{tk}}{b_{\bullet k}} ln\left(\frac{b_{tk}}{b_{\bullet k}}\right)$$

We have assumed that the overall diversity, D_γ, which is obviously $H(p)$, is a sum of within and between group diversity: $D_\gamma = D_\alpha + D_\beta$. This implies that D_β can be created via construction as $D_\gamma - D_\alpha$.

This means that D_β is

$$D_\beta = D_\gamma - D_\alpha = -\sum_t \frac{b_{t\bullet}}{V} ln\left(\frac{b_{t\bullet}}{V}\right) + \sum_k \frac{b_{\bullet k}}{V} \sum_t \frac{b_{tk}}{b_{\bullet k}} ln\left(\frac{b_{tk}}{b_{\bullet k}}\right) \tag{9}$$

This sort of by-fiat measure might seem to be arbitrary, with no reason to believe that it actually corresponds to anything meaningful. However, Appendix B demonstrates that in fact, D_β is also equal to the weighted Kullback-Leibler divergences of each group from the overall distribution. In other words, it is a measure of average distinctiveness between groups, where the score of 0 indicates that groups are identical in terms of distributions of member types.

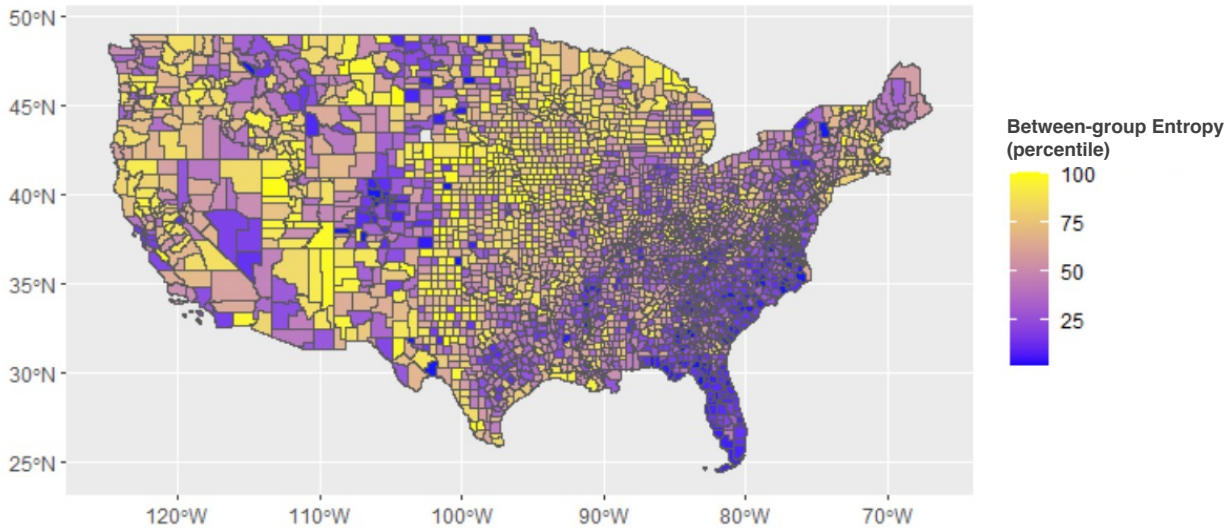
This might be of special importance to us—when people are members of groups that draw from different sections of the population. We thus consider this our final measure of social capital:

$$c^6_{\ i} = \exp\left[D_\beta\right] \tag{10}$$

When we replicate the above analyses, we do not find there to be anything remarkable in the age-by-gender distribution. But when we look at the patterns at the geographic level, we find a remarkable reversal. Some of the areas (Florida and Colorado in particular) that have high overall general diversity have extremely low diversity *between* groups (see Figure 9). All the

diversity, in other words, is coming from the fact that the groups to which they belong are internally diverse. They do not belong to *different types* of groups, if this is understood to mean drawing from different types of people. In contrast, those in the plains also tend to participate in groups that are dissimilar in their types.[5]

Figure 9: Colemanian Diversity Between-Groups by County



In sum, looking at the *types* of people (as measured by the Social Hash partition) leads us to somewhat different conclusions as to who has the most diversity in social capital. While young women seem to have the most non-redundancy when we consider the number of *concrete alters* (you know many different individuals), young men seem to have the highest non-redundancy when we consider the *types* of persons (the people you know are less likely to be similar to one another).

6. Conclusions

We hope to have not only demonstrated different ways in which we can use the Simmelian notion of the duality of groups and members to shed light on the meaning of social capital, but that we have pursued the notion of the diversity of this type of social capital in a way that both allows the use of measurement strategies well worked out in mathematical ecology, but also fits the belief that John Mohr had that "diversity" was more than a political slogan or an administrative weasel-word. The choices of (a) whether one is interested in Tocquevillian or Colemanian social capital, (b) whether to measure it in a way that accentuates its *mass* or one that accentuates *diversity,* and (c) whether to use *exogenous* or *endogenous* approaches to thinking about diversity, will of course depend on the processes of interest to the investigator. Our argument is not that one way is superior, but rather that we gain greater insight by

[5]. In all cases, the bulk of the entropy comes from within-groups as opposed to between-groups.

comparing the results of different approaches.

We have, of course, used one particular form of data, and one that has the remarkable advantage of near completeness. As shown in Table 3, some of our measures require that we have a complete census of all groups that a set of persons belong to, and all members of those groups. However, we also saw an encouraging finding that a weighted group membership measure (c^2) gave similar results to the total set of unique co-members (c^3). There may, of course, be questions or datasets (e.g., the sort of data collected by Mische 2008) for which these full-information methods may be useful.

We also examined how groups give members Colemanian social capital in the form of the *implicit* ties of co-membership. We have not investigated when and with what results individuals convert some of these implicit ties to *explicit* ties of communication and/or friendship. Doing this is a fearsome though in principle possible task, but we leave that for the future.

For the particular case of Facebook Group memberships, we have found some results that might be surprising to sociologists interested in social capital. Most important, given the assumption that "social capital," as a form of capital, is associated with other forms of privilege, and our understanding of the differences between cosmopolitans and locals, sociologists would probably expect it to be concentrated in wealthy and urban regions. Yet we have found some forms of social capital higher in less populated (c^1 high in Western mountain states) or less urban (c^2 high in the South Atlantic regions) or poorer areas (c^5 high in Appalachia as well as vacation states). It may of course be that in areas of lower population density, there is more online activity, which might suggest that online group memberships are an important form of the redistribution of social capital.

We have also found that women tend to have higher social capital than men by almost all of our measures, as they are more active in Facebook groups in general, and this ripples through almost all of our measures. And women (especially in their 30s) have more diverse Tocquevillian capital—they tend not to concentrate their attention on groups that have a high overlap of members. However, we do find that men, especially young men, have more diverse Colemanian social capital when we use the Social Hash partition to assign persons to types on the basis of their predicted probability of tie formation. While we do not have an answer to this interesting reversal, we now have a question we did not have before.

References

Adler, Paul S. and Seok-Woo Kwon. 2002. “Social Capital: Prospects for a New Concept.”

The Academy of Management Review 27(1): 17-40.

Bellah, Robert Neelly, William M. Sullivan, Richard Madsen, Ann Swidler, and Steven M. Tipton. 1985. *Habits of the Heart: Individualism and Commitment in American Life*. Berekeley: University of California Press.

Beyerlein, Kraig and John R. Hipp. 2005. “Social Capital, Too Much of a Good Thing? American Religious Traditions and Community Crime.” *Social Forces* 84 (2): 995-1013.pages)

Boggs, Carl. 2001. “Social Capital and Political Fantasy: Robert Putnam's *Bowling Alone*.”

Theory and Society 30(2): 281-297.

Bouglé, Célestin. 1926 [1922]. *The Evolution of Values*, translated by Helen Stalker Sellars. New York: Henry Holt.

Bourdieu, Pierre. 1986. “The Forms of Capital.” Pp. 241-258 in *Handbook of Theory and Research for the Sociology of Education*, edited by John G. Richardson. Westport, CT: Greenwood Press.

Breiger, Ronald L. 1974. “The Duality of Persons and Groups.” *Social Forces* 53:181-190.

Breiger, Ronald L. 2000. “A Tool Kit for Practice Theory.” *Poetics* 27: 91-115.

Breiger, Ronald L. and John W. Mohr. 2004. “Institutional Logics from the Aggregation of Organizational Networks: Operational Procedures for the Analysis of Counted Data.” *Computational & Mathematical Organization Theory* 10: 17–43

Burt, Ronald S. 1992. *Structural Holes: The Social Structure of Competition*. Cambridge, Mass.: Harvard University Press.

Burt, Ronald S. and Jennifer Merluzzi. 2016. “Network Oscillation.” *Academy of Management Discoveries* 2(4).

Carbonaro, William J. 1998. “A Little Help from My Friend's Parents: Intergenerational Closure and Educational Outcomes.” *Sociology of Education* 71:295-313

Castro, Joseph, Sarah Fenstermaker, John Mohr, and Debra Guckenheimer. 2009. “Institutional Contexts for Faculty Leadership in Diversity.” Pp. 209-230 in *Doing Diversity in Higher Education: Faculty Leaders Share Challenges and Strategies*, edited by Winnifred R. Brown-Glaude. New Brunswick: Rutgers University Press.

Chao, Anne and Chun-Huo Chiu. 2017. “Bridging the Variance and Diversity Decomposition Approaches to Beta Diversity via Similarity and Differentiation Measures.” *Methods in Ecology and Evolution* 7:919-928.

Chiu, Chun-Huo and Anne Chao. 2014. “Distance-Based Functional Diversity Measures and Their Decomposition: A Framework Based on Hill Numbers.” *PLOS One* 9(7):e00014.

Cigler, Allan and Mark R. Joslyn. 2002. “The Extensiveness of Group Membership and Social

Capital: The Impact on PoliticalTolerance Attitudes." *Political Research Quarterly* 55(1): 7-25.

Clemens, Elisabeth. 2020. *Civic Gifts: Voluntarism and the Making of the American Nation-State*. Chicago: University of Chicago Press.

Coleman, James S. 1988. "Social Capital in the Production of Human Capital." *American Journal of Sociology* 94:S95-S120.

van Dam, Alje. 2019. "Diversity and its Decomposition into Variety, Balance, and Disparity." *R. Soc. Open sci* 6: 190452. http://dx.doi.org/10.1098/rsos.190452.

Diani, Mario. 2003. "Networks and Social Movements: A Research Programme." Pp. 299-319 in *Social Movements and Networks: Relational Approaches to Collective Action*, edited by Mario Diani and Doug McAdam. New York: Oxford.

Fieldhouse, E. and D. Cutts. 2010. "Does Diversity Damage Social Capital? A Comparative Study of Neighbourhood Diversity and Social Capital in the US and Britain." *Canadian Journal of Political Science/Revue canadienne de science politique* 43(2): 289-318.

Gargiulo, Martin and Mario Benassi. 1999. "The Dark Side of Social Capital." Pp. 298-322 in *Social Capital and Liability*, edited by Leenders and Gabbay. Norwell, MA: Kluwer.

Gittell, Ross J. and Avis Vidal. 1998. *Community Organizing: Building Social Capital as a Development Strategy*. Newbury Park: Sage Publications.

Granovetter, Mark. 1973. "The Strength of Weak Ties." *American Journal of Sociology* 78:1360-1380.

Granovetter, Mark. 1985. "Economic Action and Social Structure: the Problem of Embeddedness." *American Journal of Sociology* 91(3): 481-510.

Greif, Avner. 1989. "Reputation and Coalitions in Medieval Trade: Evidence on the Maghribi Traders." *Journal of Economic History* 49: 857-82.

Hill, M. O. 1973. "Diversity and Evenness: A Unifying Notation and Its Consequences." *Ecology* 54(2):427-432.

Hooghe, Marc and Ellen Quintelier. 2013. "Do All Associations Lead to Lower Levels of Ethnocentrism? A Two-Year Longitudinal Test of the Selection and Adaptation Model." *Political Behavior* 35(2): 289-309.

Jost, Lou. 2006. "Entropy and Diversity." *Oikos* 113(2):363-375.

Jost, Lou. 2007. "Partitioning Diversity into Independent Alpha and Beta Components." *Ecology* 88(10): 2427-2439.

Kornhauser, Arthur William. 1959. *The Politics of Mass Society*. Glencoe, Ill.: The Free Press.

Kovács, Balázs. 2010. "A Generalized Model of Relational Similarity." *Social Networks* 32: 197-211.

Kurzman, Charles. 2004. *The Unthinkable Revolution in Iran*. Cambridge, Mass.: Harvard University Press.

Lande, Russell. 1996. "Statistics and Partitioning of Species Diversity, and Similarity Among Multiple Communities." *Oikos* 76: 5-13.

Li, Yaojun, Andrew Pickles and Mike Savage. 2005. "Social Capital and Social Trust in Britain." *European Sociological Review* 21(2): 109-123.

Lee, Monica and John Levi Martin. 2018. "Doorway to the Dharma of Duality." *Poetics* 68: 18-30.

Lizardo, Omar. 2018. "The Mutual Specification of Genres and Audiences: Reflective Two-Mode Centralities in Person-to-Culture Data." *Poetics* 68: 52-71.

Leydesdorff, Loet, and Ismael Rafols. 2011. "Indicators of the Interdisciplinarity of Journals: Diversity, Centrality, and Citations." *Journal of Informetrics* 5(1): 87-100.

Mantel, Nathan. 1967. "The Detection of Disease Clustering and a Generalized Regression Approach." *Cancer Research* 27(2):209–220.

Martin, John Levi. 2009. *Social Structures*. Princeton, New Jersey: Princeton University Press.

McPherson, J. Miller. 1982. "Hypernetwork Sampling: Duality and Differentiation Among Voluntary Organizations." *Social Networks* 3(4): 225-249.

Messner, Steven F., Eric Baumer, and Richard Rosenfeld. 2004. "Dimensions of Social Capital and Rates of Criminal Homicide." *American Sociological Review* 69:882-903.

Mische, Ann. 2008. *Partisan Publics*. Princeton, New Jersey: Princeton University Press.

Mohr, John W. 2000. "Introduction: Structures, Institutions, and Cultural Analysis." *Poetics* 27: 57-68.

Mohr, John and Vincent Duquenne. 1997. "The Duality of Culture and Practice: Poverty Relief in New York City, 1888-1917." *Theory and Society* 26:305-356.

Mohr, John W. and Roger Friedland. 2008. "Theorizing the Institution: Foundations, Duality and Data." *Theory and Society* 37:421-426.

Mohr, John W. and Helene K. Lee. 2000. "From Affirmative Action to Outreach: Discourse Shifts at the University of California." *Poetics* 28(1): 47-71.Oh, H., G. Labianca and M.H. Chung. 2006. "A Multilevel Model of Group Social Capital." *Academy of Management Review*, 31(3): pp.569-582.

Park, Minsu, Ingmar Weber, Mor Naaman, and Sarah Vieweg. 2015. "Understanding Musical Diversity via Online Social Media. *The 9th International AAAI Conference on Web and Social Media* (ICWSM 2015).

Portes, Alejandro and Erik Vickstrom. 2011. "Diversity, Social Capital, and Cohesion." *Annual Review of Sociology* 37: 461-479.

Preuss, Lucien G. 1980. "A Class of Statistics Based on the Information Concept." *Communications in Statistics--Theoretical and Methodological* A9(15):1563-1585.

Putnam, Robert D. 2000. *Bowling Alone*. New York: Simon and Schuster.

Reagans, Ray and Ezra W. Zuckerman. 2001. "Networks, Diversity, and Productivity: The Social Capital of Corporate R&D Teams." *Organizational Science* 12: 502-517.

Ricotta, Carlo and Laszlo Szeidl. 2009. "Diversity Partitioning of Rao's Quadratic Entropy." *Theoretical Population Biology* 76: 299-302.

Riley, Dylan John. 2010. *The Civic Foundations of Fascism in Europe: Italy, Spain, and Romania 1870-1945*. Baltimore: Johns Hopkins University Press.

Ruef, Martin and Seok-Woo Kwon. 2016. "Neighborhood Associations and Social Capital." *Social Forces* 95(1): 159-189.

Shalita, Alon, Brian Karrer, Igor Kabiljo, Arun Sharma, Alessandro Presta, Aaron Adcock, Herald Kllapi, and Michael Stumm. "Social Hash: An Assignment Framework for Optimizing Distributed Systems Operations on Social Networks." 2016. In *Proceedings of the 13th Usenix Conference on Networked Systems Design and Implementation*, pp. 455-468.

Shi, Y., Y. Lim, and C. S. Suh. 2018. "Innovation or Deviation? The Relationship Between Boundary Crossing and Audience Evaluation in the Music Field." *PloS ONE*, 13(10), e0203065. doi.org/10.1371/journal.pone.0203065.

Silver, Daniel, Monica Lee, and Clayton C. Childress. 2016. "Genre Complexes in Popular Music." *PLoS ONE* 11(5): e0155471. doi:10.1371/journal.pone.0155471.

Simmel, Georg. [1923] 1950. *Soziologie*, Third edition. Pp. 87-408 in *The Sociology of Georg Simmel*, translated and edited by Kurt H. Wolff. Glencoe, Illinois: The Free Press.

Small, Mario. 2010. *Unanticipated Gains: Origins of Network Inequality in Everyday Life*. New York: Oxford.

Stirling, Andrew. 1998. "On the Economics and Analysis of Diversity." *Science Policy Research Unit* (SPRU), ElectronicWorking Papers Series 28:1–156.

Thomson, Irene Taviss. 2005. "The Theory That Won't Die: From Mass Society to the Decline of Social Capital." *Sociological Forum* 20(3): 421-448.

de Tocqueville, Alexis. 1962 [1835]. *Democracy in America*, translated by Henry Reeve. New York: Schocken Books.

Tsallis, Constantino. 1988. "Possible Generalization of Boltzmann-Gibbs Statistics." *Journal of Statistical Physics* 52(1/2):479-487.

Velthuis, Olav. 2017. "Of Ranking and Rigging – Market Devices and Moral Economies on Chaturbate." Paper presented at the Annual Meetings of the Society for the Advancement of Socio-Economics, Lyon, June.

Viswanath, K., W. Randolph Steele, and J.R. Finnegan Jr. 2006. "Social Capital and Health: Civic Engagement, Community Size, and Recall of Health Messages." *American Journal of Public Health* 96(8): 1456-1461.

Whittaker, R. H. 1972. "Evolution and Measurement of Species Diversity." *Taxon* 21(2/3):213-251.

Wood, Gordon S. 1992. *The Radicalism of the American Revolution*. New York: Alfred A. Knopf.

Appendices

Appendix A: General Approaches to the Rao-Stirling Diversity

Here we wish to point to a flexible way of understanding the issue of joint attention used in our computation of the Rao-Stirling diversity (c^4). First, we return to the simple issue of the computation of the number group memberships. One will note that the formula given for $c^1{}_i$ may be written as $\sum \mathbf{x}_i \cdot \mathbf{x}_i$, that is, it is the sum of the dot product of itself (where the index k is implicit). We might also consider the matrix multiplication of $\mathbf{x}_i$ with itself, that is, $\mathbf{x}_i{}^t\mathbf{x}_i$. Thus if, as in our case, $\mathbf{x}_i$ is a 1×6 row vector, this operation produces a 6×6 matrix. For the first row in our table, we construct $\mathbf{x}_i{}^t\mathbf{x}_i$ as follows:

	1	*2*	*3*	*4*	*5*	*6*
1	1	1	1	0	0	0
2	1	1	1	0	0	0
3	1	1	1	0	0	0
4	0	0	0	0	0	0
5	0	0	0	0	0	0
6	0	0	0	0	0	0

We can refer to this matrix as $\mathbf{W}^i$. Any person's derived matrix assumes such a block (clique) structure. For reasons that will become clear shortly, we can also consider an alternate

$$c^{1A}{}_i = \sum \mathbf{x}_i{}^t \mathbf{x}_i = \sum \mathbf{W}^i \tag{11}$$

or a variant thereof in which we set the diagonal to zero before summing. In that case, this sum is the roundabout way of determining $c^1(c^1\text{-}1)$, the number of pairs of groups among the groups to which person A belongs. The reason to introduce this, to foreshadow, is that $\mathbf{W}^i$ may contain information on the co-activity across different sets of groups.

Let us now return to the Rao-Stirling measure, and consider generalizing so that the joint attention to two groups might be something other than an independent function of the attention given to each, as in eq. (4). Were we simply to consider a $M \times M$ matrix $\mathbf{W}^i$, the $(k,h)^{\text{th}}$ member of which indicates this joint attention, we would have an expression that is one of the large class of Mantel (1967) statistics that examine the relation between two correlations or the equivalent; examples have been well explored in network and spatial statistics. That is, we would here say

$$c^5{}_i = \sum_{k,h,\, k \neq h} w^i{}_{kh} d_{kh} \tag{12}$$

We previously defined such a $\mathbf{W}^i$ above when we had no information about group engagement other than membership, and thus eq. (11) (measure c^{1A}) is equivalent to for the special case where $\mathbf{W}^i = \mathbf{x}_i{}^t\mathbf{x}_i$ and $\mathbf{D} = \mathbf{1}$. The formula (eq. 4) for Rao-Stirling is the equivalent to $\mathbf{W}^i = \mathbf{z}_i{}^t\mathbf{z}_i$,

where $\mathbf{z}_i$ is not restricted to being the same as $\mathbf{x}_i$ (that is, we have information on *degree* of engagement). The reason to consider this wider class is both to connect to the larger family of Mantel statistics and tests, but also because in some cases, while we might indeed want to require that the contribution of any pair of groups is greatest when it is equally split between the two groups, the precise nature of the functional form is one that should be fit from the data.

Appendix B: Proof of Entropic Decomposition

By definition, D_β is

$$= -\sum_t ln\left(p_t\right) p_t + \sum_k q_k \sum_t p_{t|k} ln\left(p_{t|k}\right) \tag{13}$$

We quantify the asymmetric difference between two distributions a and b (b is the reference distribution) using the Kullback-Leibler divergence

$$KL\left(a,b\right) = -\sum_t a_t ln\left(\frac{a_t}{b_t}\right) \tag{14}$$

Note that this is an asymmetric difference; in most cases, $KL(a,b) \neq KL(b,a)$. We can rewrite D_β as follows (simply changing the order of some terms for clarity):

$$D_\beta = \sum_k q_k \sum_t p_{t|k} ln\left(p_{t|k}\right) - \sum_t ln\left(p_t\right) p_t \tag{15}$$

Now because $p_{t|k} = b_{tk} / b_{\bullet k}$ and $q_k = b_{\bullet k}/V$, we can say that $p_t = \sum_k q_k p_{t|k}$, and so are free to rewrite the last part thusly:

$$D_\beta = \sum_k q_k \sum_t p_{t|k} ln\left(p_{t|k}\right) - \sum_t ln\left(p_t\right) \sum_k q_k p_{t|k} \tag{16}$$

And then do the following manipulations:

$$D_\beta = \sum_k q_k \sum_t p_{t|k} ln\left(p_{t|k}\right) - \sum_t \sum_k q_k p_{t|k} ln\left(p_t\right) \tag{17}$$

$$= \sum_k q_k \sum_t p_{t|k} ln\left(p_{t|k}\right) - \sum_k \sum_t q_k p_{t|k} ln\left(p_t\right)$$

$$= \sum_k q_k \sum_t p_{t|k} ln\left(p_{t|k}\right) - \sum_k q_k \sum_t p_{t|k} ln\left(p_t\right)$$

$$= \sum_k q_k \sum_t \left[p_{t|k} ln\left(p_{t|k}\right) - p_{t|k} ln\left(p_t\right) \right]$$

$$= \sum_k q_k \sum_t p_{t|k} \left[ln\left(p_{t|k}\right) - ln\left(p_t\right) \right]$$

$$= \sum_k q_k \sum_t p_{t|k} ln\left(\frac{p_{t|k}}{p_t}\right)$$

$$D_{\beta} = \sum_{k} q_k KL\left(c_k, p\right).$$

In other words, D_{β} is the same as the weighted sum of the divergences of all the group entropies from the overall entropy (this is noted without comment by Lande 1996).

Van Dam (2019) shows a way to then partition D_{β} into three components: the total amount of *variety* (how many categories are there?), the *balance* (are cases clumped into a few groups, or spread out?) and the *disparity* (are the groups different?). We could consider going further in this direction, but that would take us in a different direction.

Extensions

This way of looking at exogenously defined diversity does not, like our measure for endogenously defined diversity, include the differential participation scores in **Z**. Including this information is, of course, possible, and one way to generalize would be to try to use similarities of groups in their Social Hash distributions to construct a dissimilarity matrix, to replace the **D** matrix used in the Rao-Stirling (eq. 4). In this case, the asymmetric dissimilarities that are based on information theory would be the natural equivalent to the set-based dissimilarities we created above. Thus here we would say

$$d_{kh} = \frac{H\left(\mathbf{p}_k\right) - H\left(\mathbf{p}_k \mid \mathbf{p}_h\right)}{H\left(\mathbf{p}_k\right)} \tag{18}$$

where $H()$ is the Shannon entropy (see, e.g., Preuss 1980).

However, if we have information on the degree of dissimilarities between *types*, we can follow ecological thinking and turn to a class of generalizations of the entropic diversity that includes information on dissimilarities. Appendix D demonstrates that in this case, the Rao-Stirling is a special case of this general approach, and that certain alternatives can also be decomposed into within-group and between-group diversity, but we leave the application of this to the future. But first, Appendix C establishes a useful result.

Appendix C: Demonstration of Shannon Entropy as Limit (Hill Lemma)

Here we prove what we shall call the Hill lemma (we follow Hill 1973) which is necessary for the results in Appendix D. For any probability vector **p**, let us define a generalized diversity measure of order q, ${}^{q}D$, as follows:

$${}^{q}D\left(\mathbf{p}\right) = \left[\sum_{t} p_t^{\,q}\right]^{\frac{1}{(1-q)}} \tag{19}$$

The number q is a user-chosen number that defines the "order" of the diversity measure (it is the "Hill number"). (In intuitive terms, q determines how sensitive the measure is to the presence of rare as opposed to common categories.) Note that if $q = 0$, ${}^{0}D(\mathbf{p})$ is the count of the number of types. Also note that if $q = 1$, the expression is undefined, as our exponent is 1/0. However, the

limit of ${}^{q}D(\mathbf{p})$, and that of $\ln[{}^{q}D(\mathbf{p})]$, as $q \rightarrow 1$ is defined. The Hill lemma is that the $\ln[{}^{q}D(\mathbf{p})]$, as $q \rightarrow 1 = H(\mathbf{p})$, that is, it is the Shannon entropy. First, let us define $v = q - 1$. Then

$$\lim_{q \to 1}\left(\left[\sum_t p_t{}^{q}\right]^{\frac{1}{(1-q)}}\right) = \lim_{v \to 0}\left(\left[\sum_t p_t{}^{v+1}\right]^{-\frac{1}{v}}\right) \tag{20}$$

Taking the logarithm, we find

$$\lim_{v \to 0}\left(\ln\left\{\left[\sum_t p_t{}^{v+1}\right]^{-\frac{1}{v}}\right\}\right) = \lim_{v \to 0}\left(-\frac{1}{v}\ln\left\{\sum_t p_t{}^{v+1}\right\}\right) = \lim_{v \to 0}\left(-\frac{1}{v}\ln\left\{\sum_t p_t p_t{}^{v}\right\}\right) \tag{21}$$

Since by definition,

$$p_t{}^{v} = \exp\left(v\ln\left[p_t{}^{v}\right]\right) \tag{22}$$

and making use of properties of logarithms, this is

$$\lim_{v \to 0}\left(-\frac{1}{v}\ln\left\{\sum_t p_t \exp\left(v\ln\left[p_t{}^{v}\right]\right)\right\}\right) \tag{23}$$

Now, for small values of x, $\exp(x) \approx 1 + x$ (Stirling's approximation; hence reciprocally, $x \approx \ln[1+x]$), and since $v \rightarrow 0$, these may be treated as small, so this may be written

$$\begin{aligned} {}^{q}D(\mathbf{p}) &= \lim_{v \to 0}\left(-\frac{1}{v}\ln\left\{\sum_t p_t\left(1 + v\ln\left[p_t{}^{v}\right]\right)\right\}\right) = \lim_{v \to 0}\left(-\frac{1}{v}\ln\left\{\sum_t\left(p_t + p_t v\ln\left[p_t{}^{v}\right]\right)\right\}\right) \\ &= \lim_{v \to 0}\left(-\frac{1}{v}\ln\left\{\sum_t p_t + \sum_t p_t v\ln\left[p_t{}^{v}\right]\right\}\right) = \lim_{v \to 0}\left(-\frac{1}{v}\ln\left\{1 + v\sum_t p_t\ln\left[p_t{}^{v}\right]\right\}\right) \end{aligned} \tag{24}$$

By the reciprocal version of Stirling's approximation, this can be written

$${}^{q}D(\mathbf{p}) = \lim_{v \to 0}\left(-\frac{1}{v}v\sum_t p_t\ln\left[p_t\right]\right) = \lim_{v \to 0}\left(-\sum_t p_t\ln\left[p_t\right]\right) = H(\mathbf{p}) \tag{25}$$

This will be of use to us below.

Appendix D: Demonstration of Relations Between Diversities and Entropish Measures at Different Orders

We begin by following Ricotta and Szeidl (2006) in sketching an interpretation of the Shannon entropy and generalizing to a wider variety of entropish functions. (Some use the term *entropy*

to mean any of these generalized weighted surprisal functions; for the purposes of clarity we only use the word to refer to the Shannon/Boltzman entropy.) Recall that the Shannon entropy is

$$H(\mathbf{p}) = -\sum_t p_t ln(p_t) \tag{26}$$

This may be understood as containing two portions, the *rarity* of any type, and its *surprisal*. These may seem to be the same thing (for we are surprised to find a very rare type), but there are good information-theoretic reasons for saying that, for type t, the rarity is tapped by p_t and the surprisal by $-\ln(p_t)$. We will be generalizing by changing the surprisal function based on what we know about the similarity of certain types.

Let us begin by noting that, obviously, $p_t = 1 - \sum_{u \neq t} p_u$, hence we may say

$$H(\mathbf{p}) = -\sum_t p_t ln(p_t) = -\sum_t p_t ln\left(1 - \sum_{u \neq t} p_u\right) \tag{27}$$

This surprisal implies that all types are equally foreign to one another. But if type u^* were fundamentally the same as type t, we probably would not want to count its prevalence as part of the surprisal for type t; hence we can imagine a generalization of equation 27

$$H(\mathbf{D},\mathbf{p}) = -\sum_t p_t ln(p_t) = -\sum_t p_t ln\left(1 - \sum_{u \neq t} d_{tu} p_u\right) \tag{28}$$

where $\mathbf{D} = \{d_{tu}\}$ is a matrix of dissimilarities between types (please do not confuse this dissimilarity matrix **D**, always in Roman bold, with the diversity functions D, always in italics—we were running out of letters).

It is this intuition, coupled with the Hill lemma (see Appendix C), that allows a generalization. First, note that $\sum_{u \neq t} d_{tu} p_u$,is the t^{th} row of the compound matrix **Dp**. With this as our conceptual guideline, let us instead of a matrix of differences, begin with a matrix of similarities **S**, to mesh the derivation with that of our sources.

Following Leinster and Cobbold (2012), and with **p** defined as previously, and **S** a $T \times T$ matrix of similarities between types (however determined), such that $0 \leq s_{tu} \leq 1$, $s_{tt} = 1$; consider the generalization of eq. 19

$${}^qD(\mathbf{p},\mathbf{S}) = \left[\sum_t p_t \left([\mathbf{Sp}]_t\right)^{q-1}\right]^{\frac{1}{(1-q)}} \tag{29}$$

where q is the Hill number. In this notation, the ${}^qD(\mathbf{p})$ of eq. 19 may be written ${}^qD(\mathbf{p}, \mathbf{1})$. Note that $\mathbf{Sp}_t$, that is, the t^{th} row of **Sp**, is the sum of the similarities of the t^{th} type to all others; if the **S** matrix is suitably normalized, this is the expected similarity between this type and an individual of another type chosen at random. Let us first consider the case in which **S** is defined as a diagonal **1** matrix (that is, all values on the diagonal are 1, all off-diagonal elements are 0). Then $\mathbf{Sp}_t = s_{t1}p_1 + s_{t2}p_2 + \ldots + s_{tt}p_t \ldots + s_{tT}p_T = s_{tt}p_t = p_t$. In this case, we find eq. 19 reprinted here for clarity

$$ {}^{q}D(\mathbf{p},\mathbf{1}) = \left[\sum_t p_t^{\,q} \right]^{\frac{1}{(1-q)}} \tag{30} $$

By Hill's lemma, we have seen that the limit of this expression as $q \rightarrow 1$ is $\exp(H(\mathbf{p}))$. Now consider the case when $q = 2$. This becomes

$$ {}^{2}D(\mathbf{p},\mathbf{1}) = \left[\sum_t p_t^{\,2} \right]^{-\frac{1}{1}} = \frac{1}{\sum_t p_t^{\,2}} \tag{31} $$

which is equivalent to the inverse of the Herfindahl measure of concentration, and is the diversity reached by the Gini-Simpson index, the Renyi entropy, and others (Jost 2006: 364).

Now let us return to the more general case in which **S** is empirically observed similarities across types. First, when $q = 1$, it is not hard to use Hill's lemma to show that

$$ {}^{1}D(\mathbf{p},\mathbf{S}) = \lim_{q \to 1} \left(\left[\sum_t p_t \left([\mathbf{Sp}]_t \right)^{q-1} \right]^{\frac{1}{(1-q)}} \right) = \exp\left[-\sum_t p_t \ln\left(\sum_{u \neq t} s_{tu} p_u \right) \right] = \exp\left[H(\mathbf{D},\mathbf{p}) \right] \tag{32} $$

as defined above (indeed, Hill first defined his system using more general weights).

Now consider the case where $q = 2$ (but **S** is not necessarily **1**):

$$ {}^{2}D(\mathbf{p},\mathbf{S}) = \left[\sum_t p_t \left([\mathbf{Sp}]_t \right) \right]^{-1} = \frac{1}{\sum_{t,u} p_t s_{tu} p_u} \tag{33} $$

Using this notation, let us return to the entropies. We have already seen that the Shannon entropy is indeed simply the logarithm of ${}^{1}D(\mathbf{p},\mathbf{1})$. But let us consider Tsallis's (1988) generalization of the continuous Boltzmann entropy,

$$ H^* = k \frac{1 - \sum_t p_t^{\,q}}{q-1} \tag{34} $$

where k is some constant. While in many physical applications, we would set k to Boltzmann's constant, since here any positive constant is acceptable, we choose $k = 1$ for simplicity. Tsallis shows that this has the properties needed of a measure (additivity, invariance under transformations). Again, the limit of H^* $q \rightarrow 1$ is the Shannon H, and again, let us generalize this to accept similarities **S**

$$ {}^{q}H(\mathbf{p},\mathbf{S}) = \frac{1 - \sum_t p_t \left[\mathbf{Sp} \right]_t^{q-1}}{q-1} \tag{35} $$

This might seem unlikely to turn out to be a reasonable "entropy." But a key identity of the natural logarithm is that

$$\lim_{\delta\to 0}\left(\frac{x^{\delta}-1}{\delta}\right)=\ln(x) \tag{36}$$

which allows for the following generalization where we define the following family of functions

$$\ln_q(x)=\frac{x^{1-q}-1}{1-q} \tag{37}$$

By setting $\delta = 1 - q$ we can see that $ln_q(x) \rightarrow ln(x)$ as $q \rightarrow 1$. Now it can be seen that

$$\ln_q\left({}^qD(\mathbf{p},\mathbf{Z})\right)=\frac{\left\{\left[\sum_t p_t\left([\mathbf{Sp}]_t\right)^{q-1}\right]^{\frac{1}{(1-q)}}\right\}^{1-q}-1}{1-q}=\frac{\left[\sum_t p_t\left([\mathbf{Sp}]_t\right)^{q-1}\right]-1}{1-q}=\frac{1-\sum_t p_t\left([\mathbf{Sp}]_t\right)^{q-1}}{q-1} \tag{38}$$

Now since

$${}^qD(\mathbf{p},\mathbf{Z})^{1-q}=\left\{\left[\sum_t p_t\left([\mathbf{Sp}]_t\right)^{q-1}\right]^{\frac{1}{(1-q)}}\right\}^{1-q}=\sum_t p_t\left([\mathbf{Sp}]_t\right)^{q-1} \tag{39}$$

for cases where $q \neq 1$, by equation 37, the corresponding entropy at this order is given by

$${}^qH(\mathbf{p},\mathbf{S})=\ln_q\left({}^qD(\mathbf{p},\mathbf{S})\right)=\left(\frac{1}{q-1}\right)\left(1-{}^qD(\mathbf{p},\mathbf{S})^{q-1}\right) \tag{40}$$

So for the case $q = 2$

$${}^2H(\mathbf{p},\mathbf{S})=\left(\frac{1}{2-1}\right)\left(1-{}^2D(\mathbf{p},\mathbf{S})^{1-2}\right)=\left(1-\left[\frac{1}{\sum_{t,u}p_t s_{tu}p_u}\right]^{-1}\right)=1-\sum_{t,u}p_t s_{tu}p_u \tag{41}$$

We have been working in terms of similarities **S**, but given the range of *s*, we can convert this into dissimilarities **D**, that is, $d_{tu} = 1 - s_{tu}$. In that case

$${}^2H(\mathbf{p},\mathbf{S})=1-\sum_{t,u}p_t s_{tu}p_u=1-\sum_{t,u}p_t p_u\left(1-d_{tu}\right)=1-\sum_{t,u}p_t p_u+\sum_{t,u}p_t p_u d_{tu}=\sum_{t,u}p_t p_u d_{tu} \tag{42}$$

(since the sum of all products of a probability vector with itself to any power is 1). In other words, the entropy of the second order is found to be the Rao-Stirling diversity.