The Geography of Facebook Groups in the United States

Amaç Herdağdelen[1], **Lada Adamic**, **Bogdan State**[2]

[1] Core Data Science, Meta
[2] Scie.nz
amac@herdagdelen.com, ladamic@gmail.com, bogdan@scie.nz

Abstract

We present a de-identified and aggregated dataset based on geographical patterns of Facebook Groups usage and demonstrate its association with measures of social capital. The dataset is aggregated at United States county level. Established spatial measures of social capital are known to vary across US counties. Their availability and recency depends on running costly surveys. We examine to what extent a dataset based on usage patterns of Facebook Groups, which can be generated at regular intervals, could be used as a partial proxy by capturing local online associations. We identify four main latent factors that distinguish Facebook group engagement by county, obtained by exploratory factor analysis. The first captures small and private groups, dense with friendship connections. The second captures very local and small groups. The third captures non-local, large, public groups, with more age mixing. The fourth captures partially local groups of medium to large size. Only two of these factors, the first and third, correlate with offline community level social capital measures, while the second and fourth do not. Together and individually, the factors are predictive of offline social capital measures, even controlling for various demographic attributes of the counties. To our knowledge this is the first systematic test of the association between offline regional social capital and patterns of online community engagement in the same regions. By making the dataset available to the research community, we hope to contribute to the ongoing studies in social capital.

Introduction

Understanding the norms and bonds that allow communities to act together constitutes one of the key focus areas for modern social science. The concept of *social capital* has long been used to refer to a collection of measures at the individual, inter-personal and community level. These measures account for the strength of a diffuse fabric of trust relations, group affiliations and organizational structures. Social capital allows individuals to draw benefit from their connections, while also enabling communities to act collectively for common benefit (Scott and Johnson 2005). The degree to which social capital varies between countries and communities has elicited a great deal of interest. In the United States in particular, the extent to which social capital measures vary between counties is a well-established sociological finding.

The rise of Internet-based social networking has facilitated the development of place-based virtual communities. Multiple platforms – community forums, local subreddits, local Facebook pages, groups and events, Neighborly pages, to name but a few – have emerged that cater to the need of virtual connection in local communities. Their importance has been bolstered by difficulty of in-person interaction during the COVID-19 pandemic. The development of such virtual arenas raises an interesting question as to the extent to which known differences in offline social capital are reproduced in the online world.

Facebook – the world's largest social network at the time of writing and the focus of our analysis – is an important platform facilitating connections within local communities, in particular through the medium of the "Facebook Groups" product. Facebook groups support a wide variety of local communities, for people bound together by shared neighborhood issues, hobbies, interests, or affiliations. These include groups corresponding to local organizations, such as neighborhood associations, scout troops, and sports clubs, to more informal associations such as those between parents of a cohort within a school district or a local meetup group for various hobbies and interests. Still other local groups have a commercial focus, such as "for sale" groups, while others support gift exchange as part of the "buy nothing" movement.

It is not surprising to note that many Facebook groups' existence, membership, and interactions are shaped by offline contexts. For example, the number, size, and characteristics of Facebook groups may in part depend on the presence of offline organizations. Furthermore, interaction norms in the offline context may also be present online. This observation brings forth the expectation of community-level differences in social capital also being present in the online realm of local Facebook groups in particular, given the strong influence that local contexts are expected to have on these virtual entities.

The question of how to operationalize and measure social capital remains an open one. Measures of social capital include *generalized trust* (defined as the extent to which individuals believe unknown alters can be trusted; see Bjørnskov, 2007), as well as summary indices of civic and social participation. Because of the difficulty involved in collecting large-scale measures of social capital, few existing datasets lend themselves to a systematic analysis at the local level. A particularly notable exception comes from the Social Capital Project of the Senate's Joint Economic Committee (Social Capital Project 2018), which estimated county-level social capital measures across multiple dimensions.

We examine the extent to which the geographic patterns revealed by this dataset are also found in online interactions facilitated by local Facebook groups. Our analysis focuses on "on-platform" indicators: 42 aggregated and de-identified county-level metrics of Facebook Group usage. These include, among others, the share of users in the county who are in small/medium/large and private/public groups, proportion of private groups observed in the county, whether admins require membership or post approval, etc. We perform exploratory factor analysis on these measures. We then examine the correlations between the computed factors and offline social capital estimates. Finally, we investigate the relationship between the online group factors and measures of community-level public health and inequality outcomes which are mediated by social capital.

Related Work

Prior work has examined the positive association between individuals' activity on online social networking sites (SNS) and social capital (Ellison, Steinfield, and Lampe 2007; Vanden Abeele 2018; Steinfield, Ellison, and Lampe 2008; Vanden Abeele 2018; Tiwari, Lane, and Alam 2019), though increases in social capital depend on the type of use (Burke, Kraut, and Marlow 2011). These studies focused on perceived social capital at the individual level, by asking participants whether they had online contacts they could turn to for various needs.

At the community level, however, studies investigating the connection between online interactions and social capital have typically been limited to individual localized communities, such as a college (Ellison, Steinfield, and Lampe 2006) or the "Netville" neighborhood that was given broadband internet access earlier than others (Wellman, Boase, and Chen 2002). This line of work revealed important conclusions applicable at the individual level. For instance, local online interactions were found to be associated with knowing more individuals in one's community and having higher bridging social capital. Our work aims to shift the focus from individual-level relationships to entire communities. Doing so is only possible when correlating local social capital measures with online activity across a large geographic scale, a possibility afforded to us by the examination of Facebook groups across the United States.

Social capital has come to refer to a particularly immediate understanding of the institutional milieu, a broad set of factors including but not limited to: interpersonal connections, reciprocity, trust and trustworthiness, and shared norms and identity (Lochner, Kawachi, and Kennedy 1999). Higher social capital allows communities to mobilize effectively in the face of crises and makes it possible to solve collective action problems. Such communities fare better in a multitude of societal outcomes such as health, safety, and resilience. In Putnam's characterization, social capital greases the wheels that allow communities to advance smoothly (Putnam et al. 2000, p. 288).

The exact meso-level mechanism through which the loose set of concepts termed to be social capital impacts collective outcomes remains an open area of debate. Transaction costs economics (Williamson 2008) holds that societies with high social capital have lower transaction costs in everyday business and social life. A differing view is held by sociological neo-institutionalists (Granovetter 1985), who identify in social capital a deeper set of shared understandings which is not directly reducible to interactional cost accounting. Regardless of its exact mechanism of action, there is ample evidence of differences in social capital across countries, as evidenced for instance in comparative studies of international datasets such as the World Values Survey (Minkov and Hofstede 2012; Bjørnskov 2007).

Trust is an important aspect to the functioning of online communities as well as social capital in communities in general. Diversity within a geographic community was shown by Putnam to be associated with lower trust between groups and within groups in a large survey conducted in 2000 (Putnam 2007). Generally, people place more trust in smaller groups (La Macchia et al. 2016). In the online setting, Ma et al. (2019) found people to place more trust in Facebook groups which were smaller, private as opposed to public, having denser networks of friendships within the group, as well as greater age and gender homogeneity. These findings were also supported by interviews of participants in "Mom-to-mom" for-sale groups on Facebook, who built trust in private groups of similar members and active group admin involvement. Trust in groups correlated with individuals' propensity to trust and feelings of receiving social support from others in general. Interestingly, increasing participation corresponded to subsequent increases in feelings of trust for the group, suggesting that online engagement may contribute to building social capital (Iyer et al. 2020).

Some characteristics of online friendship ties, when aggregated to the county level, have been found to correlate with offline social capital. For example, Bailey et al. (2018) found counties with a higher proportion of Facebook ties to friends living more than 100 miles away have higher measures of social capital, and other variables such as income, high school graduation, life expectancy, and social mobility. However, community-level associations, such as the structure and participation in online groups, may be more directly informative about the formation and activation of *community* social capital in a location. This is the subject of our present study.

Data and Methods

We examine the association between community-level social capital by analyzing a dataset that summarizes the activity of US Facebook groups at county level. This data set is available online[1] (Herdagdelen, Adamic, and State 2022).

[1] https://doi.org/10.7910/DVN/OYQVEP

For the purpose of this analysis, we consider a group to be active if it has (1) at least one comment, like, or posting event in the 28 days preceding the cutoff date, and (2) at least 10 active members. We consider a member to be active if they had taken at least one action (commenting, posting, liking, etc.) in the group in the 28 days preceding the analysis date. When we compute county-level aggregate statistics over group memberships, we only consider active members. We consider a group a "US group" if at least 90% of its members are in the US. For the aggregated age and gender distribution metrics we use self-reported values by the users. Location data is based on predicted city of the users. These values may contain inaccuracies due to self-reports and prediction errors. This is a limitation of our measurements.

Defining Group Locality

We take a pragmatic approach and use the intuition that a "local group" is composed of people who live near one another, regardless of its function or utility.

A "local" Facebook group may be directly tied to a local organization (e.g., a school group), a physical place or institution (e.g., friends of a library), interest group (neighborhood watch, local for sale groups), or just happen to bring people together living in the same area. We consider the mechanism through which a group is salient to a particular location beyond the scope of our study.

As an aggregate measure of locality of a group g, we use the probability of two randomly-chosen members of the group being in the same county c, or $\lambda_g = P(C_i = C_j)$ where C_i (C_j) is a random variable indicating the county associated with the member i (j) of the group g, with $i \neq j$. This locality measure also known as Simpson index is computed as:

$$\lambda_g = P(C_i = C_j) = \frac{\sum_c n_c(n_c - 1)}{N(N-1)},$$

where n_c is the number of members of g in county c, while N is the total number of members of group g. At one extreme, if all members of a group live in the same county the measure is 1. If each member lives in a different county, the measure would take its minimum value 0. If half of the members live in the same county and all other members live in isolated counties, the Simpson Index of the group would converge to 0.25 in asymptotically as N grows.

As a validation of the locality measure, we looked at probabilities of observing various uni- and bi-grams in public and non-hidden private group names, as a function of locality measure. In Figure 1, we see the relation between the observation probabilities of four illustrative n-grams.

Groups that mention "family reunion" tend to be non-local, hinting at the fact that families that reunite tend to live far apart. Groups that mention "parents" in the name tend to be very local. "Yard sale" groups are more likely have mid-level locality scores suggesting that these kinds of buy and sell groups are of intermediate locality, spanning several counties. The n-gram "support group" illustrates the bimodal nature of support groups; they are more frequently either very local, or very non-local.

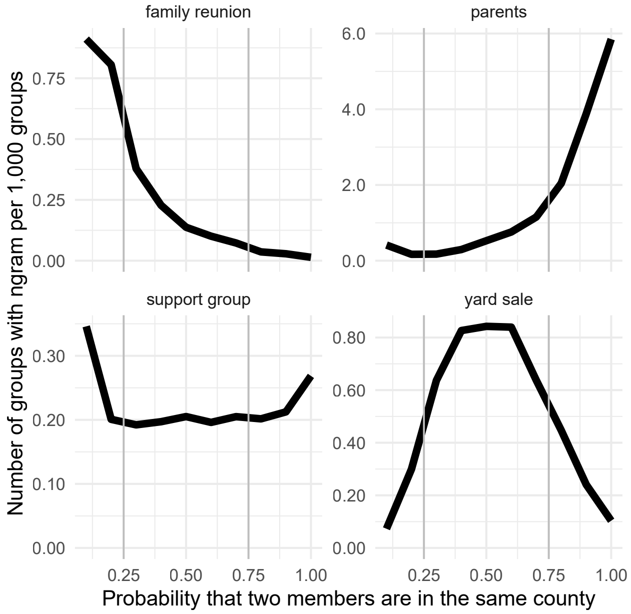


Figure 1: Observation probabilities of illustrative n-grams as a function of group locality score.

Informed by this anecdotal evidence, we decided to set two locality score thresholds: 0.25 and 0.75, dividing groups into three buckets. A group is called *very local* if the probability of two random members being in the same county is above 0.75, *local* if the same probability is between 0.25 and 0.75, and *non-local* if the estimated probability is lower than 0.25. With these thresholds, any group that contains a majority of their members in one county gets classified as local, and if an overwhelming majority lives in the same county it's classified as very local, allowing us to differentiate between groups that are local but still attract memberships from neighboring counties and groups that are strictly localized in one county.

Membership in Local Groups When aggregating local group membership to the county level, we consider the county of the member, even if the group is local to a different county. For example, if a user living in county A is a member of a group that has over 50% of its members in county B, this would still count as a "local" membership for county A. This broader definition captures participation that is still limited to groups where a majority of members belong to the same county, but allows a group to have a more flexible local "area of influence", even if it extends across sometimes arbitrary-seeming county borders.

Group Size

We consider two group-level characteristics to have particular relevance to the relationship between local Facebook groups and community-level social capital. We expect group dynamics to vary as group size increases and group members are less likely to be directly connected through social relations. Given the extent to which trust is constitutive of social capital, we likewise expect important interactions between group privacy levels and social capital.

We partitioned groups according to their membership size m into four buckets: $m < 30$ is very small, $30 \leq m < 100$ is small, $100 \leq m < 1000$ is mid-size, and groups with at least 1,000 members are categorized as large. The strategy of dividing groups into progressively-larger size buckets was chosen due to the heavy-tailed nature of the group size distribution.

Group Privacy

Facebook groups have different privacy levels that determine the who can see the group content and membership list. *Public* groups have the highest visibility and both the content and member list are visible to non-members. Content in *Private* groups is only visible to members. Furthermore, private groups can be *hidden*, meaning that the groups itself is only visible to its members and can be joined by invitation.

Offline Social Capital Estimates

Efforts to measure social capital systematically and at scale have been comparatively rare. This dearth of information is particularly pronounced for additional requirements of recency and geographic details – both necessities for providing a detailed picture of the current state of social capital in the United States.

The Social Capital Project (2018) of the Senate's Joint Economic Committee (JEC) offers the most detailed and timely estimate of social capital we were able to identify at the time of the study. In 2018, this project released a comprehensive dataset of multiple social, economic and physical indicators, aggregated at the level of individual US counties. Some of these indicators are used as components of a composite social capital estimate and some are used as benchmarks against which the social capital estimate is evaluated in terms of predictive power. Given our theoretical interest in associational social capital, we focus on the *community health* components of the Social Capital Project.

Community Health is a composite index based on the number of county-level non-religious non-profits per person, religious congregations per person, and state-level data on percentage of people who attended public meetings, report working with neighbors to solve common problems, participate in volunteer work, etc. This is one subdimension used in the JEC study, but since it captures the associational activity in a county, we decided to use it as our dependent variable.

As we see in Figure 2, counties with higher population tend to have lower values of the community health index, with an overall correlation of (-0.58). While the implications of this relation are beyond the scope of this study, interested readers can refer to the literature on scaling effects of populations and organizations and resources (Gastner and Newman 2006; Bettencourt et al. 2007).

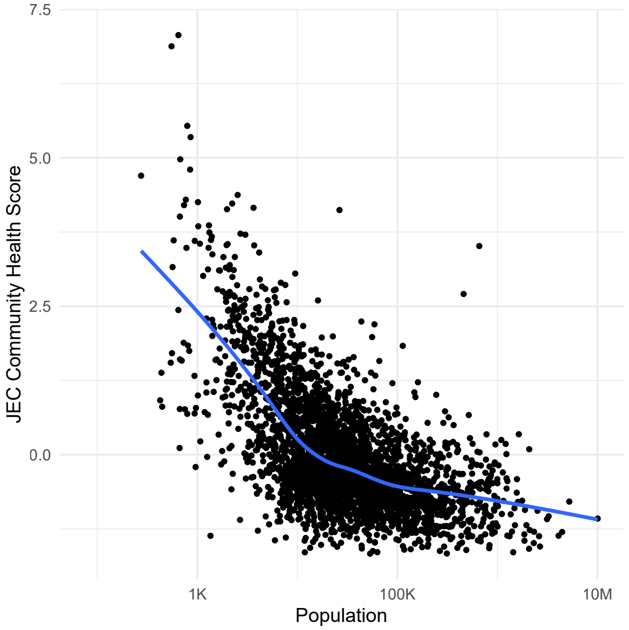


Figure 2: Community Health Indices of US counties, as a function of county population. Spearman correlation between the two is -0.58.

We also compute a community health index (y_i) adjusted for population and density by fitting an OLS with the logarithm of these values (and their squares) as the only covariates and using the residuals as the adjusted community health index, $y_i = \beta_1 log(p) + \beta_2 log(p)^2 + \beta_3 log(d) + \beta_4 log(d)^2 + \epsilon_i$, where p and d are total population and population density of county i, respectively, with y_i the community health index of county i. We use the residual ϵ_i as the adjusted index. We will use both unadjusted and adjusted indices in our exploratory analyses.

Generalized Trust (occasionally called "social trust") is one of the more common operationalizations of social capital (Newton 2001). Broadly speaking, it is defined as the willingness to trust a generic alter, or "the potential readiness of citizens to cooperate with each other and [the] abstract preparedness to engage in civic endeavors with each other" (Stolle 2002). Generalized trust is typically measured attitudinally, at the individual level – the relevance of this indicator being bolstered by the inclusion of a generalized trust question in the World Values Survey. The work of Bjørnskov (2007) reveals striking country-level differences for this construct. In the United States, strong inter-regional differences were documented by Simpson (2006). Recently, Wu (2020) even established the persistence of discrepancies in generalized trust attitudes for individuals moving between regions with varying levels of generalized trust.

Generalized trust is arguably uniquely suited as an individual-level construct that can nonetheless reveal deep, community-level differences in social capital. Percentage of people living in an area who exhibit high levels of trust is used as a community-level indicator. Unfortunately, a county-level generalized trust dataset against which one could compare this indicator is lacking.[2]. State-level gen-

[2] This gap in existing measurement was also identified by the authors of the afore-mentioned JEC study (Social Capital Project 2018, p.19)

eralized trust measurements are however available thanks to the Generalized Social Survey (Neville 2012). We rely on these measures as the best available indicators of social trust for local geographies in the United States.

Group-Based Indicators

We compute 42 aggregated and de-identified county-level indicators of Facebook Groups usage. These indicators can be grouped under two categories:

Participation in Different Group Types In this category we measure what percentage of active Facebook users in a county participate in different group types by size, privacy, and locality. First, we partition each group into one of 36 mutually exclusive buckets, organized along three dimensions based on privacy (public / private / hidden), number of users in the group (very small, small, mid-sized, large), and locality (very local, local, non-local). Then we compute the percentage of users living in a county who participate actively as a member in at least one such group. For instance, the indicator (private, very local, very small) for a county represents the percentage of Facebook users who live in the county and who are active members of at least one very small, private group with a very high locality score.

Aggregate Local Group Characteristics Here we aggregate the characteristics of local groups observed in a county. We only use groups where the majority of members live in the same county (i.e., locality score $\geq$ 0.25) and include a group in the averaging only for the county where the majority of members live in[3]. The characteristics are:

- *Content gating*. Ratio of local groups in a county where admin approval is required for posts before they become visible to other members.
- *Member gating*. Ratio of local groups in a county where the group has at least one of three controls for joining as a new member: admin approval, agreeing to group rules, or questions that are required to be answered for joining.
- *Mean gender diversity*. We use the probability that two randomly chosen members of a group having different genders as the gender diversity of the group. This measure is the complement of Simpson's index for gender distribution (also known as the Blau score). Formally, gender diversity of a given group g is $\gamma_g = 1 - P(G_i = G_j)$ where G_i (G_j) is a random variable indicating the self-reported gender (male or female) of a member i (j) of the group g, with $i \neq j$.

  The average Blau score of local groups observed in a county c is used as the county-level gender diversity indicator for the county:

  $$\frac{\sum_{g \in G_c} \gamma_g}{|G_c|},$$

  where G_c is the set of all local groups in county c, and γ_g is the group-level gender diversity as defined above.
- *Mean locale diversity*. We use the probability that two randomly chosen members of a group having different locale setting for Facebook as the locale diversity of the group. Example locales include "English (US)" or "Portugês (Brasil)." Average locale diversity scores that are local to a county is used as the mean locale diversity indicator for the county. Formally, this is defined in a similar manner to the gender diversity defined above,

  $$\frac{\sum_{g \in G_c} \mu_g}{|G_c|},$$

  where $\mu_g = 1 - P(M_i = M_j)$, giving us the probability that locale settings of two randomly chosen members of g are different.
- *Mean age diversity*. Average inter-quartile range of the age distribution of local groups. The inter-quartile range is defined as the number of years between the 25th and 75th percentile values of the empirical age distribution.
- *Mean tie density*. Ratio of actually realized (made) Facebook friendship ties between group members to all possible pairs.

High colinearity among our indicators renders standard regression techniques inappropriate for exploratory data analysis. We could collapse these indicators, using marginal participation rates along salient dimensions, instead of the joint measurements along privacy, size, and locality. However, in the absence of a strong theoretical basis and *a priori* expectations of which indicators go together, we decided to employ exploratory factor analysis to study the internal correlation structure of online platform indicators and their relation to offline societal benchmark indicators.

Results

Before we discuss the results in more detail, we illustrate three on-platform indicators to build a stronger intuition of the data we are studying.

1. Local Group Membership Prevalence is defined as the ratio of users who live in a county and who are active members of at least one local or very local group, to all active members in the same county. In the following exploratory analyses, we use sub-components of this indicator, broken down by privacy, size, and locality of the groups, but for illustration purposes here we use the composite indicator.

A county-level choropleth (Figure 3) reveals strong spatial autocorrelation and large-scale patterns in local group membership prevalence. In certain areas of the Midwest, Great Plains, Coastal California, and particularly the Deep South, we observe low levels of local group membership prevalence.

Intuitively, local group membership prevalence may be expected to reflect associational activity and social capital in an area. However, we do not observe a correlation (ρ = 0.04) between *% local users* and *community health*, the JEC index measuring associational activity in a county, such as volunteering, attending public meetings, working with neighbors to fix things, etc.

[3] We also used a less restricted way of incorporating a group in averaging for a county, using the ratio of group members in the county as a weight. The result did not change meaningfully.

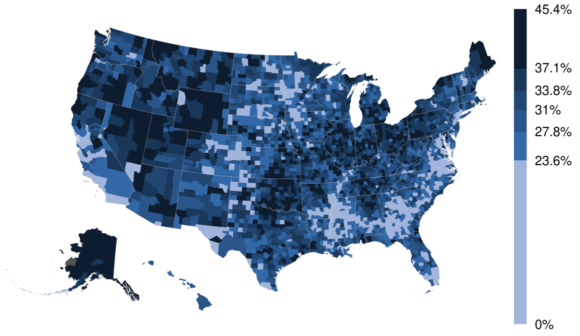


Figure 3: Share of active local group members (%) who are in at least one local group all active Facebook users. Each color represents an equal-sized (in terms of counties) bucket.

2. Content Control The second example indicator is *p(content gated)*, the percentage of local groups that do not allow members to post content without admin approval, shown in Figure 4. The choropleth reveals a gradient of increasing content control from North to South. The lowest levels of the indicator are observed in parts of the Midwestern United States (North Dakota, South Dakota, Minnesota, Iowa, and Wisconsin). The highest levels of the indicator are observed in the Southern States (Louisiana, Mississippi, Alabama, Georgia, South Carolina, Texas, and Florida).

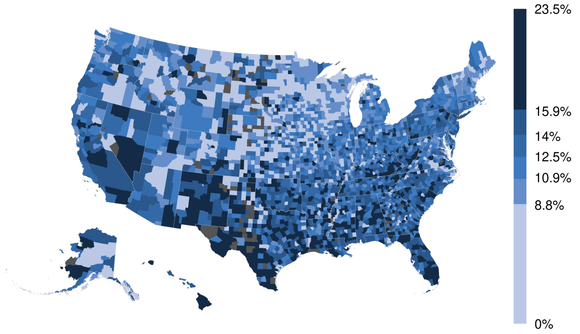


Figure 4: Percentage of groups in each county that require admin approval for contents posted to the group.

We aggregated the data at the state level to represent what % of local groups require admin approval. Doing so allowed us to compare our findings against the previously-discussed trust questions in the General Social Survey (Neville 2012). We observe a strong negative relation ($\rho = -0.81$) between the share of groups with admin approval and the % of survey respondents who agreed that strangers can be trusted. *p(content gated)* also has similarly strong correlation ($\rho = -0.82$) to the trust in neighbors measure provided by the Social Capital Project (2018). A scatter plot of both measures depicting the strong bivariate relation is given in Figure 5.

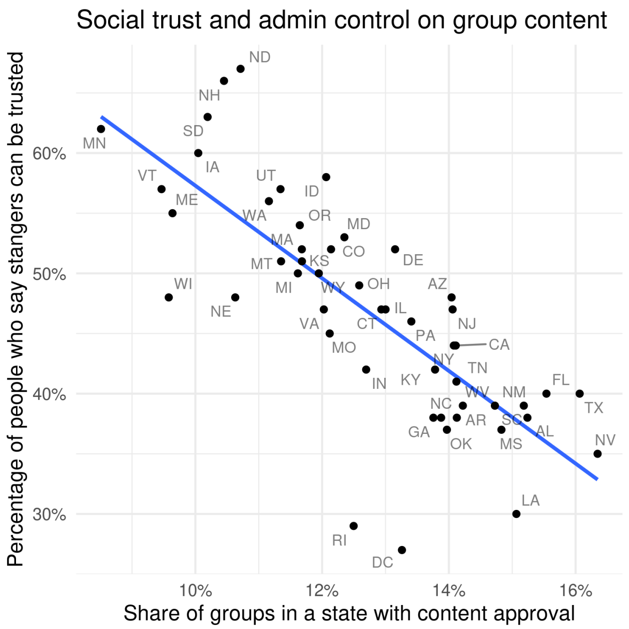


Figure 5: Generalized trust (in strangers) and share of groups *without* admin approval are strongly correlated ($r = 0.81$).

However, such correspondence is not shown by all group-derived indicators that one might hypothesize reflect trust. For example *p(member gated)*, the probability that a group places controls on who can become a member of the group, is moderately correlated with trust in neighbors (-0.56), but exhibits only a weak correlation with generalized trust (-0.25).

While one could come up with plausible *post-hoc* explanations for why the above results make sense, our conclusion is not to trust our hypotheses about single indicators, despite their plausibility. Instead, we use a data-driven approach to unearth latent factors that all of our indicators measure collectively. In the next section, we provide the results of exploratory factor analysis.

3. Linguistic Diversity Figure 6 of the above-defined *m*ean locale diversity reveals an unsurprising pattern: Areas that receive in-migration, such as border counties, the West Coast, and other urban areas, have higher levels of locale diversity, reflecting the demographic characteristics of these areas. In Figure 7, we plot the locale diversity of the groups in each county with respect to the percentage of foreign-born population and note a strong correlation ($\rho = 0.61$).

Extracting Latent Factors

To identify the latent factors that our indicators can collectively explain, we applied exploratory factor analysis (EFA). We used the maximum likelihood method and extracted 4 factors as suggested by a visual inspection of the scree plot (see Online Appendix[4] for details), using VARIMAX rotation. The factor structure was robust to using OBLIMIN rotation and we obtained qualitatively similar results using a varimax-rotate principal component analysis. Factor loadings are summarized in Figure11, while loading values are visualized for each county in the Appendix, which also pro-

[4] https://arxiv.org/abs/2211.09043

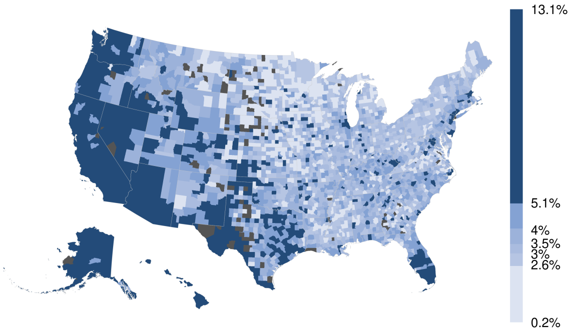


Figure 6: Average locale diversity of local groups.

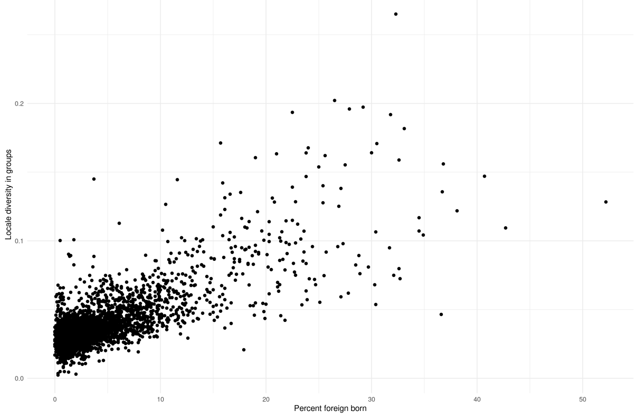


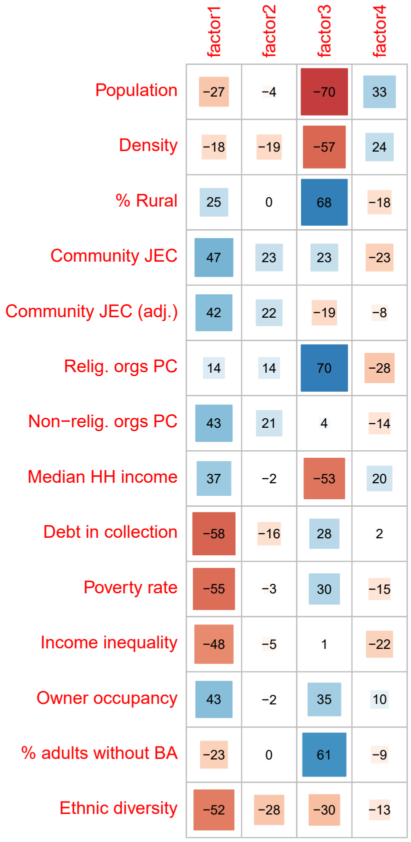
Figure 7: Locale diversity and percentage foreign-born population in each county ($\rho = 0.61$.)

vides explicit factor loadings for all the component variables.

We flipped the signs of the factors in a way that made interpretations easier. The sign selection is essentially an arbitrary decision that does not change numerical results or their interpretation.

In the following sections, we follow the same recipe to describe each factor in order. First, we interpret the factor loadings on individual indicators and note common patterns. Then, we report correlations with population and community health measures. We also take note of relevant socioeconomic and demographic indicators which likewise exhibit relatively strong correlations. We also present and discuss the geographical patterns over the choropleths of the factors and note pockets of high- and low-value areas.

Finally, we provide a list of group name ngrams that have high salience for each factor: Using the regression factor loadings from the county-level dataset, we compute a "faux-factor" score for each group (e.g., if a group is a public, local, small group we use the factor loading of public_local_small). Then, we compute the average group factor scores of unigrams and bigrams occurring in group names.



	factor1	factor2	factor3	factor4
Population	−27	−4	−70	33
Density	−18	−19	−57	24
% Rural	25	0	68	−18
Community JEC	47	23	23	−23
Community JEC (adj.)	42	22	−19	−8
Relig. orgs PC	14	14	70	−28
Non−relig. orgs PC	43	21	4	−14
Median HH income	37	−2	−53	20
Debt in collection	−58	−16	28	2
Poverty rate	−55	−3	30	−15
Income inequality	−48	−5	1	−22
Owner occupancy	43	−2	35	10
% adults without BA	−23	0	61	−9
Ethnic diversity	−52	−28	−30	−13

Figure 8: Spearman correlation between the factors and social indicators.

The top-scoring ngrams for each factor are identified as salient ngrams.

F1: Small, Tightly Knit, Non-local Groups F1 captures counties with small non-local groups with dense ties, low content and member gating, and low locale diversity, In Fig. 11, we plot the loading factors of group-participation indicators. This factor's values are highest in the Midwest and Utah, and lowest in California and the Southwest (Figure 10).

Figure 8 reveals that F1 is correlated to both unadjusted ($\rho = 0.47$) and population-adjusted community health ($\rho = 0.42$). The geographical distribution is visibly similar to that of community health (see Fig. 9). Among the two major components of JEC's measure, F1 is correlated with non-religious, non-profit organizations per capita ($\rho = 0.43$), but not with religious organizations per capita ($\rho = 0.14$).

F1 is also positively correlated with economic health indicators, namely, median household income ($\rho = 0.37$) and owner-occupied housing ($\rho = 0.43$). In addition, F1 is negatively correlated with debt in collection ($\rho = -0.58$, poverty rate ($\rho = -0.55$) and income inequality ($\rho = -0.48$). Finally, it is negatively correlated with ethnic diversity in the county ($\rho = -0.52$).

A detailed analysis of the relationship between diversity and social capital is beyond the scope of our study. How-

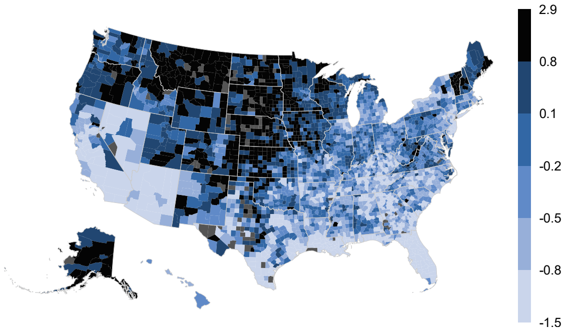


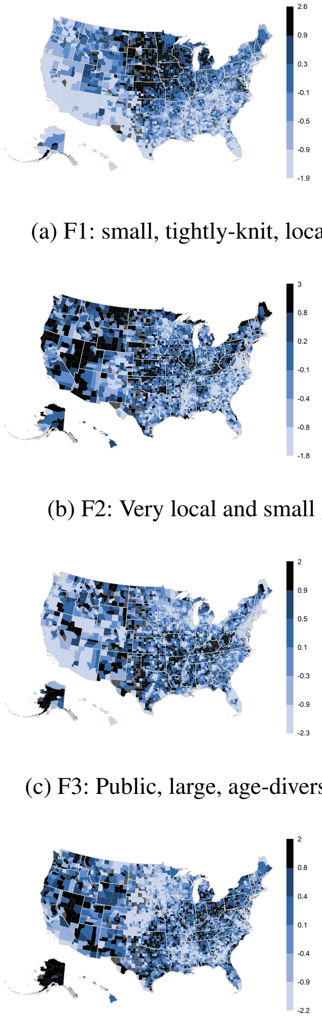
Figure 9: Choropleth of JEC community health subindex.

ever, given the negative associations between F1 and admin control, locale diversity of groups, and ethnic diversity of the county populations, we wanted to provide an additional layer of analysis, inspired by the literature on trust and diversity (Simpson 2006). In Table 1, we provide the coefficient estimates of a linear model with F1 as the dependent variable, focusing on the coefficient estimates of the binary indicator of whether the county is in the Southern United States and ethnic diversity, while controlling for population, percent rural, median household income, high school graduation rate, and religious congregations per capita. The reported coefficient estimates are for standardized input variables (mean subtracted and divided by two standard deviations, following Gelman (2008)).

The main effect sizes of *is_south* (-0.071) and *ethnic diversity* (-0.539) are in line with the correlations we reported above and also previous discussions of diversity and trust and social capital (Putnam 2007). Higher ethnic diversity and being in the Southern US are associated with lower F1 values. However, the high value of the coefficient of the interaction term *is_south:ethnic_diversity* (0.818) suggests that in the Southern United States, the relation between diversity and F1 goes against the direction predicted by trust literature. In the Southern US, counties with more ethnic diversity exhibit higher levels of F1. The interaction term effectively cancels out the negative contribution of increasing ethnic diversity in the South.

In Table 2, the salient ngrams for F1 support our interpretation of this factor. Family-related group ngrams such as “cousins,” “clans,” “family group,” “descendants” explain the tight-knit, but also geographically non-local nature of this factor. We speculate that in areas with high F1, geographically separated family groups are over-indexed.

Surprisingly, we also observe many multi-level marketing (MLM) related ngrams such as “usborne books,” “pampered chef,” “zyia,” etc. MLM is an industry that relies on a nonsalaried workforce making direct sales to customers. These nonsalaried sellers (also called consultants, distributors, etc.) take commissions out of their sales and also try to grow the seller network by recruiting downline consultants and taking a cut of these downline sales. MLM activity is known to rely on existing social capital and consultants need to tap a variety of bonding and bridging connections in their efforts to sell the products (Lofthouse and Storr 2021).



(a) F1: small, tightly-knit, local groups.

(b) F2: Very local and small groups.

(c) F3: Public, large, age-diverse groups.

(d) F4: Partially local, member-controlled and bridging groups.

Figure 10: Choropleths of factor values at the county level.

While a closer analysis of the MLM phenomenon on Facebook groups is outside of the scope of this paper, the relation between MLMs and F1 is so strong that we were concerned that F1 could simply be capturing MLM activity and related groups, without further ecological validity. As a robustness check, we went back and repeated our analyses from end to end, excluding groups that contain popular MLM brands and related keywords in their names (See Appendix for a full list of these keywords). The factor analysis results proved to be stable and still allowed us to extract a factor that is over-indexed in the Midwest (see Appendix).

These results, taken together with the robustness check,

	Coef.	*S.E.*
log(population)	-0.440^{***}	(0.041)
% rural	0.072^{**}	(0.036)
median household income	0.387^{***}	(0.029)
% adults graduated high school	0.776^{***}	(0.032)
religious congregations per 1000	0.273^{***}	(0.035)
is south	-0.071^{**}	(0.028)
ethnic diversity	-0.539^{***}	(0.028)
is south : ethnic diversity	0.818^{***}	(0.049)
Constant	-0.076^{***}	(0.012)
Observations	3,027	
R^2	0.617	
Adjusted R^2	0.616	
Residual Std. Error	0.599 (df = 3018)	
F Statistic	606.672^{***} (df = 8; 3018)	

*p<0.1; **p<0.05; ***p<0.01

Table 1: Coefficient estimates of a linear model estimating F1. Numerical variables are standardized by centering around 0 and divided by 2 standard deviations.

are consistent with the idea that MLM activity relies on the existing social capital stock of communities. In communities with higher trust and stronger bonds, people might find it easier to market the products and recruit downline consultants. The non-local nature of F1 is also consistent with MLM activity. We suspect that Facebook groups might be helping MLM participants to tap into their geographically long-distance ties to avoid hyper-local competition.

F2: Very Local and Small Groups The indicators that best measure F2 are the ratio of users who are in (very) local and (very) small groups (Figure 11). Fig. 8 further reveals F2 is not strongly correlated with population ($\rho = -0.04$), nor with density ($\rho = -0.19$) or rural population ($\rho = 0$), and is very weakly correlated with community health ($\rho = 0.23$) and its population-adjusted version ($\rho = 0.21$). We do not observe any substantial correlation between F2 and any of the demographic or economic benchmark indicators ($\rho < 0.30$ for all), but F2 is slightly negatively correlated county ethnic diversity ($\rho = -0.28$).

Geographical pockets of the U.S. exhibit particularly low values of F2. New Mexico and Colorado, most of the South (Mississippi, Georgia, North and South Carolina, Virginia), Minnesota, and the San Francisco Bay Area in California, have low values of F2 (Figure 10).

Table 2 shows a clear pattern of ngrams related to very local, neighborhood- and family-oriented groups: "girl scout" troops, "neighborhood watch," "HOA," (shorthand for Homeowners' Associations) "kindergarten," "preschool," "soccer," "mr" and "mrs" (we suspect these titles are related to teachers of classes), etc.

It is somewhat surprising that a factor that seems to capture the prevalence of very local, neighborly group participation seems to exhibit virtually no correlation with social capital measures and a very weak correlation with external estimates of social capital and community health.

	F1	F2	F3	F4
1	cousins	girl scout	paparazzi	homeschool
2	clan	scout troop	auto	singles
3	family group	troop	creations	arizona
4	spring nail	scout	crafts	church of
5	s family	kindergarten	custom	houston
6	s usborne	mrs	parts	areas
7	trip	grade	rocks	surrounding areas
8	family	20 21	loving memory	county
9	nail party	2020 2021	used	and surrounding
10	extended	scouts	in loving	baptist church
11	nail bar	pack	memory of	surrounding
12	family page	4 h	found	barter
13	usborne	preschool	cars	items
14	birthday	estates	baptist church	yardsale
15	spring into	girl	small business	uncensored
16	books more	neighborhood	car	buy nothing
17	usborne books	4th	who like	church
18	descendants	subdivision	friends who	garage sale
19	descendants of	neighbors	in memory	volunteers
20	zyia party	3rd	services	clothes
21	s spring	hoa	sale or	yard sale
22	virtual pampered	2nd	memory	online yard
23	bash	book club	or trade	pokemon
24	chef party	committee	buy sale	garage
25	into	2020	for sale	yard

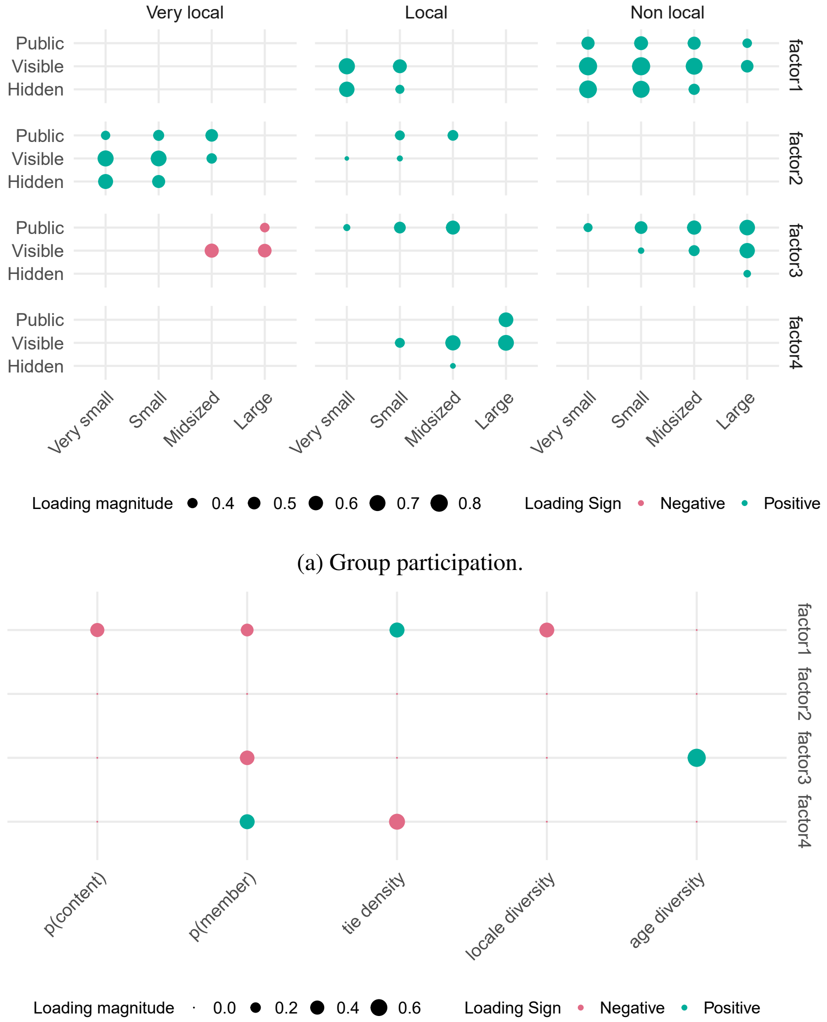
Table 2: Group name ngrams with the highest average factor scores for each factor. See text for methodology. Only ngrams that are observed in at least 500 groups are included.

F3: Public, Age-diverse Groups with 100+ Members Indicators that contribute to F3 include age diversity (as captured by the average inter-quartile range of the members' age distribution in local groups) and the share of large and public groups and memberships (Fig. 11).

One could interpret the openness of groups, inter-age mixing, and participation in large groups, with access to many other individuals, as positively correlated with bridging social capital. That said, F3 is only weakly correlated with population-unadjusted community health (0.22). F3 also correlates positively with sparsely populated rural areas ($\rho = 0.67$), which tend to have higher community health, and negatively with population and density. In fact, the correlation F3 has with community health changes signs once we adjust for population ($\rho = -0.19$). Thus, we believe F3 captures the contrast between the Facebook Groups usage patterns in small towns vs. urban centers. Population and rural population seem to be the mediator that drives the relation between this factor and community health. F3 is also strongly correlated with religious organizations per capita ($\rho = 0.70$) and percentage of adults without a bachelor's degree ($\rho = -0.61$).

We speculate that small towns may allow everyone to participate in fairly big, well-mixed settings. On the other hand, urban areas, where large, inclusive groups may grow too large for many purposes, could support smaller, more exclusive, and age-differentiated groups.

The geographical distribution of F3 (Figure 10) strength-



(a) Group participation.

(b) Local group characteristics.

Figure 11: Visual representation of indicator loadings on factors

ens our interpretation: the value of this factor is elevated in parts of the Great Plains, Appalachia, and is depressed on the West Coast, North East Coast, and northern parts of the Midwest, including large urban centers.

A small puzzle is the contrasting result of positive loading of the percentage of people in private non-local and large groups (0.701) and negative loading of the percentage of users who are in private, very local, and large groups (0.592). Large and very local groups, by definition, require many members who live in the same county and this is possible in areas with a high population density. Large groups in rural areas, by necessity, span multiple counties and end up with low locality scores. This is a good reminder of the potential effects of our operationalization of locality which may have different meanings in rural and urban areas.

The n-gram analysis in Table 2, reveals a patchy pattern: ngrams related to transactional groups (buy and sell, services, used, etc.), ngrams that might indicate memorial groups (loving memory, in memory), and again some MLM-related keywords ("paparazzi", an MLM brand, crafts).

F4: Partially Local, Member-controlled and Bridging Groups All group participation indicators (all but tie density and member gating) load on this factor with a positive weight (Figure 11). The indicators span different privacy and size buckets, but all of them are limited to local group types (but not non-local or very local groups). Especially midsized, large and local group participation indicators load on F4. In addition, this factor captures counties where groups have low friendship density among the members (loading -0.584) and do not involve membership gating procedures (loading 0.505). The findings suggest the factor is capturing partially local large group activity bringing together people not already bound by friendship ties.

F4 has a slight correlation with population ($\rho = -0.33$) and a weak negative correlation with percentage of rural population ($\rho = -0.18$). The weak negative correlation between F4 and community health ($\rho = -0.23$) disappears when controlling for population ($\rho = 0.08$). We don't observe any notable correlation with any of the other social benchmark indicators – which, given this factor might be capturing local group activity, is notable on its own.

When we look at the most salient ngrams associated with F4, we see terms related to local associations (homeschool, church, volunteers) and terms that indicate a locality at large scales (surrounding areas) along with terms related to transactional groups (buy and sell, yardsale) and neighborhood groups that foster local gift economies (buy nothing).

Multiple regions score particularly low on F4: 1) Midwest and Great Plains, 2) Southern counties, 3) 1) parts of California, except most of Central Sierra Nevada, 4) parts of Virginia and North Carolina (Figure 10). Low F4 scores in these regions indicate relative lack of memberships in local groups with sparse friendship ties that cover multiple counties.

Discussion

Our results reveal per-county geographic variation within the United States in participation in Facebook groups of various sizes, visibility, locality, age mixing, friendship density, and membership and content gating. We show that a substantial portion of the variability across all these dimensions can be represented by four main dimensions.

F1, corresponding to small, private and non-localized groups, is one of the two factors correlated with offline social capital. It is highest in the Midwest, and its choropleth is visually similar to that of the JEC social capital index. This correspondence could be due to higher levels of social capital allowing for the formation of such groups, as friends can invite one another and sustain activity in the group without it necessarily needing to be local or open to all. An example of small groups relying on non-local friendship ties is that of multi-level marketing, though the results hold with such groups excluded. Furthermore, since counties with higher F1 have more non-religious organizations per capita, some group activity likely corresponds to participation in those formal organizations. F1 is also negatively correlated with the poverty rate.

F2, representing participation in small or local Facebook groups, was not strongly correlated with offline indicators. This lack of association was also consistent with our initial exploration of participation in any local groups. This was a surprising result, if one expects areas with high social capital to have more interactions between neighbors. However, it may be that very local groups do not draw particularly on social capital, but rather depend on associations that may occur almost by default, e.g. families whose kids play on the same soccer team or are in the scouts, residents living in a neighborhood or building with a homeowners' association, etc.

F3 captures interactions in more sparsely populated areas, where separation by age, or limiting participation by making a group private, makes less sense. F3 is negatively correlated with county population, median income, and % of adults with a college degree, while being positively correlated with % rural, and the number of religious organizations per capita. The positive association between F3 and the offline social capital indices reverses direction when we control for population and density of the counties.

F4 highlights areas where people who are not already friends interact in partly local groups. That F4 is not correlated with offline social capital indices is again somewhat surprising, given that some of the groups it captures correspond to community organizations and activities such as volunteering, church, and homeschooling. Then again, it also captures transactional groups, such as for-sale groups, where social capital may not play as much of a role.

Although we have demonstrated strong correlation of online factors and informative grouping with respect to offline social capital indicators, we note that our analysis is based on differences in collective behavior on one platform, Facebook, on one product, Groups. Thus, it may contain biases due to differential adoption of Facebook and participation in groups by age and region. Other online platforms that people use as a substitute to Facebook Groups may introduce similar biases. In this sense, we have limited data, but by providing the dataset to the research community we hope that future research will incorporate other sources and will be able see a more complete picture than we could at present.

Conclusion and Future Work

In this paper we presented the first analysis of the correspondence between online group-based interaction in a given location, and social capital measures of the same. The results are correlational. Most likely the offline environment is what we see reflected in online groups. On the other hand, it is possible that some communities are able to use social media, including Facebook, to foster greater social capital. For example, perhaps social media is able to support the kind of private, small groups that are more typical of rural areas, but for people living in dense areas. Similarly, finding new shared interests can overcome the lower trust that diverse communities have (Putnam 2007). The incremental value that social media provides to social capital could be the subject of future work.

Another area of future exploration could utilize online indicators to detect changes in community social capital by region. More careful calibration with offline data, incorporating differential adoption, would be needed before one could potentially use Facebook variables to estimate changes in offline social capital. However, such indices could indeed be produced continuously, and only periodically calibrated with more expensive and time intensive surveys and offline data collection.

Given the promising results in this first analysis, we are sharing the Facebook group interaction dataset in the hopes that it may be incorporated in other studies of county-level differences, whether in measuring social capital or other inquiries, such as online collective action, civic engagement, or disaster response.

Ethics Statement

While the research was carried out, the authors were either employees or contractors of Meta Inc., the owner of the Facebook social networking platform, which also provided the funding and resources for this work. The principal benefit of the data and analysis is uniquely rich insight into the deeper structures of social capital, especially as they unfold on increasingly-important online platforms. As detailed in the literature review, social capital is a crucial factor for the health of local communities. The most important risk identified relates to breaches of user privacy. To minimize this risk we (1) look at user location information coarsened to the county level and (2) aggregate all computed indicators across all groups local to a county. No individual-level data was used in the study and the dataset is limited to counties with at least 100 users contributing to the aggregates. The authors confirm having read the AAAI Code of Ethics and Conduct and their commitment to abide by it.

References

Bailey, M.; Cao, R.; Kuchler, T.; Stroebel, J.; and Wong, A. 2018. Social connectedness: Measurement, determinants, and effects. *Journal of Economic Perspectives*, 32(3): 259–80.

Bettencourt, L. M.; Lobo, J.; Helbing, D.; Kühnert, C.; and West, G. B. 2007. Growth, innovation, scaling, and the pace of life in cities. *PNAS*, 104(17): 7301–7306.

Bjørnskov, C. 2007. Determinants of generalized trust: A cross-country comparison. *Public choice*, 130(1): 1–21.

Burke, M.; Kraut, R.; and Marlow, C. 2011. Social capital on Facebook: Differentiating uses and users. In *CHI'11*, 571–580.

Ellison, N.; Steinfield, C.; and Lampe, C. 2006. Spatially bounded online social networks and social capital. *International Communication Association*, 36(1-37).

Ellison, N. B.; Steinfield, C.; and Lampe, C. 2007. The benefits of Facebook "friends:" Social capital and college students' use of online social network sites. *Journal of computer-mediated communication*, 12(4): 1143–1168.

Gastner, M. T.; and Newman, M. E. 2006. Optimal design of spatial distribution networks. *Physical Review E*, 74(1): 016117.

Gelman, A. 2008. Scaling regression inputs by dividing by two standard deviations. *Statistics in medicine*, 27(15): 2865–2873.

Granovetter, M. 1985. Economic action and social structure: The problem of embeddedness. *American journal of sociology*, 91(3): 481–510.

Herdagdelen, A.; Adamic, L.; and State, B. 2022. Replication Data for: The Geography of Facebook Groups in the United States, https://doi.org/10.7910/DVN/OYQVEP. *Harvard Dataverse*.

Iyer, S.; Cheng, J.; Brown, N.; and Wang, X. 2020. When Does Trust in Online Social Groups Grow? In *ICWSM*, volume 14, 283–293.

La Macchia, S. T.; Louis, W. R.; Hornsey, M. J.; and Leonardelli, G. J. 2016. In small we trust: Lay theories about small and large groups. *Personality and Social Psychology Bulletin*, 42(10): 1321–1334.

Lochner, K.; Kawachi, I.; and Kennedy, B. P. 1999. Social capital: a guide to its measurement. *Health & place*, 5(4): 259–270.

Lofthouse, J. K.; and Storr, V. H. 2021. Institutions, the social capital structure, and multilevel marketing companies. *Journal of Institutional Economics*, 17(1): 53–70.

Ma, X.; Cheng, J.; Iyer, S.; and Naaman, M. 2019. When Do People Trust Their Social Groups? In *CHI'19*, 1–12.

Minkov, M.; and Hofstede, G. 2012. Hofstede's fifth dimension: New evidence from the World Values Survey. *Journal of cross-cultural psychology*, 43(1): 3–14.

Neville, L. 2012. Do economic equality and generalized trust inhibit academic dishonesty? Evidence from state-level search-engine queries. *Psychological Science*, 23(4): 339–345.

Newton, K. 2001. Trust, social capital, civil society, and democracy. *International political science review*, 22(2): 201–214.

Putnam, R. D. 2007. E Pluribus Unum: Diversity and Community in the 21st Century. *Scandinavian Political Studies*, 30(2): 137–174.

Putnam, R. D.; et al. 2000. *Bowling alone: The collapse and revival of American community*. Simon and schuster.

Scott, J. K.; and Johnson, T. G. 2005. Bowling alone but online together: Social capital in e-communities. *Community Development*, 36(1): 9–27.

Simpson, B. 2006. The poverty of trust in the Southern United States. *Social Forces*, 84(3): 1625–1638.

Social Capital Project. 2018. The Geography of Social Capital in America. Technical Report 1-18, Joint Economic Committee - Republicans.

Steinfield, C.; Ellison, N. B.; and Lampe, C. 2008. Social capital, self-esteem, and use of online social network sites: A longitudinal analysis. *Journal of Applied Developmental Psychology*, 29(6): 434–445.

Stolle, D. 2002. Trusting strangers–the concept of generalized trust in perspective. *Österreichische Zeitschrift für Politikwissenschaft*, 31(4): 397–412.

Tiwari, S.; Lane, M.; and Alam, K. 2019. Do social networking sites build and maintain social capital online in rural communities? *Journal of rural studies*, 66: 1–10.

Vanden Abeele, e. a. 2018. Does Facebook use predict college students' social capital? *Communication Studies*, 69(3): 272–282.

Wellman, B.; Boase, J.; and Chen, W. 2002. The networked nature of community: Online and offline. *It & Society*, 1(1): 151–165.

Williamson, O. E. 2008. Transaction cost economics. In *Handbook of new institutional economics*, 41–65. Springer.

Wu, C. 2020. Does migration affect trust? Internal migration and the stability of trust among americans. *The Sociological Quarterly*, 61(3): 523–543.