EPJ Data Science
a SpringerOpen Journal

REGULAR ARTICLE

Open Access

CrossMark

Is social capital associated with synchronization in human communication? An analysis of Italian call records and measures of civic engagement

Marco Mamei[1], Francesca Pancotto[1], Marco De Nadai[2,3*] , Bruno Lepri[3], Michele Vescovi[4], Franco Zambonelli[1] and Alex Pentland[5]

*Correspondence: denadai@fbk.eu
[2] University of Trento, Povo, Italy
[3] Fondazione Bruno Kessler, Povo, Italy
Full list of author information is available at the end of the article

Abstract

Social capital has been studied in economics, sociology and political science as one of the key elements that promote the development of modern societies. It can be defined as the source of capital that facilitates cooperation through shared social norms. In this work, we investigate whether and to what extent synchronization aspects of mobile communication patterns are associated with social capital metrics. Interestingly, our results show that our synchronization-based approach well correlates with existing social capital metrics (i.e., Referendum turnout, Blood donations, and Association density), being also able to characterize the different role played by high synchronization within a close proximity-based community and high synchronization among different communities. Hence, the proposed approach can provide timely, effective analysis at a limited cost over a large territory.

Keywords: Social capital; Mobile phone data; Computational social science

1 Introduction

Synchronization is a process that allows the automatic coordination of units and events in time. Across many domains in nature, it is a mechanism that permits to reduce uncertainty and risk without the need for a centralized mechanism of control. Synchronization is a widespread phenomenon observed everywhere in nature, from animals [1] to neurons [2] and heart cells [3], and up to more complex entities like human beings [4, 5].

In humans, synchronization emerges as a spontaneous coordination mechanism that provides benefits to groups and the individuals that live within [6]. In an evolutionary perspective, synchronization increases the probability of group survival, by reducing the individual costs required by the engagement of coordinated and cooperative action [7]: in a multilevel selection mechanism, a group of cooperators has indeed higher chances of evolutionary success than a group of defectors. The positive effect of synchronization is also found in the behavior of people within groups, where synchronous activity has been found to enhance the level of cooperativeness [8] even without *muscular bonding* [9] or *shared positive emotions* [10, 11]. Synchronized groups should then in principle be more

cooperative ones, and by comparing the level of synchronization between different groups, we may be able to measure their relative level of cooperativeness. In the present study, we propose two *synchronization indices*: (i) *within synchronization* representing the relative level of cooperation within a close proximity-based community (i.e., municipality level), and (ii) *between synchronization* representing the level of cooperation among different communities in a larger geographical area (i.e., province level). More specifically, these indices capture the synchronization of human activity in an area through mobile phone data. Mobile phone data capture rich information about human activities and the structure of the social interactions therein [12]. They have been used to estimate the socioeconomic status of territories [13] and individuals [14], to analyze the dynamics of cities [15], to model the spreading of diseases [16], and to predict crime levels [17]. Our hypothesis is that the two synchronization indices, capturing the degree of cooperativeness among human activities, can describe traditional measures of social capital, which is the source of capital that facilitates cooperation through shared social norms [18].

The relevance of social capital for economic growth is largely acknowledged [19]; it reduces the transaction costs associated with formal coordination mechanisms, [20] predicts strong economic performance [21] and financial development [22], and reduces corruption by inducing political and civic participation [23, 24].

An important distinction in the social capital literature is the one between *bonding* and *bridging* patterns of relations [25]. In his work, the political scientist Putnam states that *bonding social capital* provides emotional support and a sense of belonging in which the members of a community sustain each other [25]. This form of social capital is usually observed in homogeneous groups with strong cooperation, such as families or circles of close friends. *Bridging social capital*, instead, stems from relations between groups, that is, between individuals from heterogeneous backgrounds [25]. A community exploring novel interactions and co-operation with other communities can be considered to have a high amount of bridging social capital [26]. This form of social capital has been described as potentially useful for achieving instrumental goals since a larger variety of resources becomes available by interacting with people of diverse status, occupation or ethnicity [26].

Previous research on capturing bonding and bridging social capital, and their effect on economic prosperity, from mobile phone and social media data has analyzed this issue focusing on the role played by different network structural properties (e.g., topological network diversity, network density, etc.) [13, 27]. To the best of our knowledge, the current work is the first study that analyzes whether and to what extent synchronization aspects of human communication are associated with traditional social capital metrics (i.e., Referendum turnout, Blood donations, and Association density).

Several studies have highlighted the role and the benefits played by the synchronization of activities among individuals and groups. Indeed, synchronization is argued to improve cooperation and trust in a community [5, 8]. Hence, we expect that communities with strong synchronization may experience richer opportunities for cooperation, decreased costs of market interactions, less reliance on formal business regulations and increased informal money circulation and investments, all aspects enabled by high levels of trust [5, 8, 28]. Thus, our *first hypothesis* is that high levels of call activity's synchronization in a tight area (that we associate to a municipality) are likely to reflect *bonding* patterns as people interact and communicate within a close proximity-based social group. In par-

ticular, high levels of *within synchronization* in a proximity-based community capture frequent communication patterns and connections among people living in this community.

Interaction among diverse groups of individuals and communities have been linked to higher exploration of possibilities, thus promoting the flow of information and novel ideas that affect economic prosperity [2, 6]. Following Paxton [29], *bridging* social capital occurs when members of one group connect with members of other groups to seek access, support or to gain information. On this basis, our *second hypothesis* is that the interaction of a given community (i.e., a given municipality) with many different communities can be found in the high synchronization of their communication patterns. In particular, we expect that municipalities with more synchronization with other municipalities may experience a communication with a more diverse array of communities (i.e., having bridging ties spreading to many different municipalities) and gain novel ideas and information, and thus may show higher levels of *bridging* social capital.
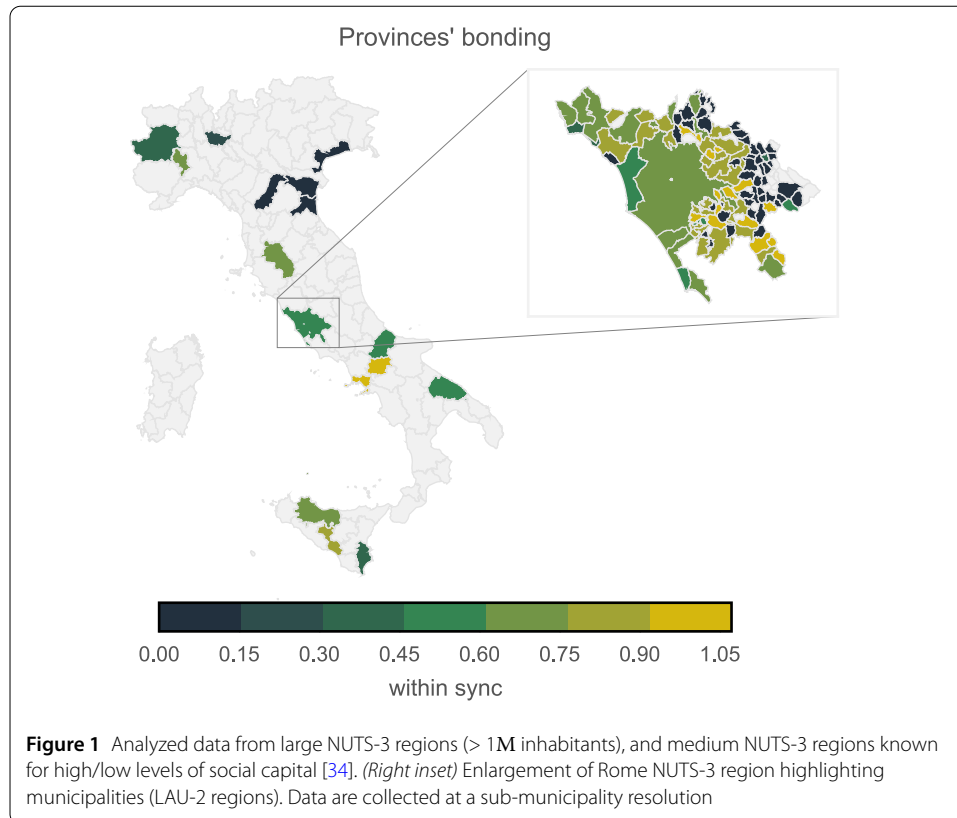
Interestingly, our results show that a synchronization-based approach well correlates with traditional social capital measures (i.e., Referendum turnout, Blood donations, and Association density), being also able to characterize the different role played by high synchronization within a close proximity-based community and high synchronization among different communities.

2 Materials and methods

For this study we use an aggregated and anonymized Call Detail Records (CDRs) dataset provided by the largest Italian mobile phone operator (34% of market share) over a period of one month: from March 31, 2015 to April 30, 2015. CDRs are collected for billing purposes by mobile network operators: every time a phone interacts with the network, a CDR recording the time and location (in terms of cell network's antenna) of the user is created.[a] The data we use is spatially aggregated and completely anonymized by the mobile phone operator as it is not possible to connect different calls of the same user.

Italy is an ideal playground in this domain because Italian regions present very different levels of economic development, although they have experienced the same formal institutions, laws, language and currency for many years now. Many scholars have identified the root of this persistent divergence in differential endowments of social capital [30, 31]. For these reasons, Italy has been widely studied in social capital economic literature [23, 25]. As a byproduct, there are several survey-based data sources for obtaining social capital measures that can be used as a ground-truth. More specifically, following examples in the economics literature [22, 25, 32], we use *Referendums turnout*, *Association density* and *Blood donations* as our ground-truth. *Referendums turnout* are usually considered as proxy of the desire of civic participation, as voting at referendums is not mandatory in Italy and the issues on the ballot in referendums are less related to local interests. *Association density* is defined as the number of associations per 100,000 inhabitants. Associations can be cultural, leisure, artistic, sports, environmental, and any kind of nonprofit associations with the exclusion of professional and religious associations [19]. *Blood donations* are measured as the instances of donations per 1000 inhabitants.

In our analysis, we select both large provinces (NUTS-3 regions) with more than one million inhabitants, and smaller provinces known for high and low levels of social capital (according to the aforementioned social capital survey-based measures). The indicators

Figure 1 Analyzed data from large NUTS-3 regions (> 1M inhabitants), and medium NUTS-3 regions known for high/low levels of social capital [34]. *(Right inset)* Enlargement of Rome NUTS-3 region highlighting municipalities (LAU-2 regions). Data are collected at a sub-municipality resolution

of level of social capital used to select small NUTS-3 regions—intended with a population between 200,000 and 500,000 inhabitants—are the data available for Italy on association density, referendum participation and blood donations [30, 33, 34]. Specifically, considered NUTS-3 regions are:
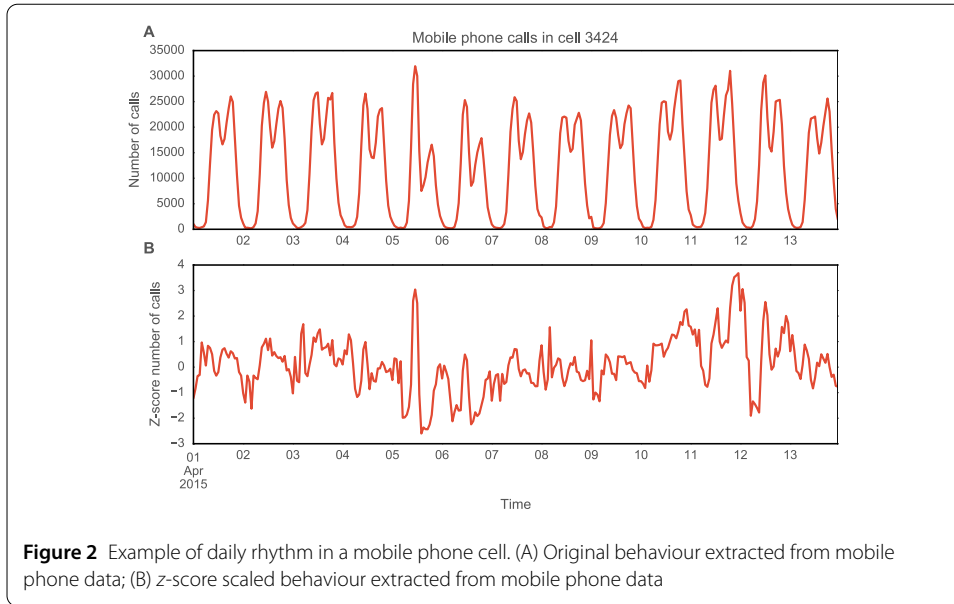
- Turin, Milan, Venice, Rome, Naples, Bari, Palermo (large NUTS-3 regions);
- Caltanissetta, Siracusa, Benevento, Campobasso (defined as low-social capital NUTS-3 regions [34]);
- Siena, Ravenna, Ferrara, Asti, Modena (defined as high-social capital NUTS-3 regions [34]).

These areas represent the smallest areal units available for social capital data. NUTS-3 regions are therefore our unit of analysis. The choice of these NUTS-3 regions is partly data-driven, but we select them also as they exhibit different levels of social capital. Figure 1 shows the map of Italy with the NUTS-3 regions under analysis.

The area of each region is spatially divided in an irregular grid, provided by the mobile phone operator, based on the size of the underlying antennas' coverage area. The cells have area ranging from 0.04 km^2 in the city center to 40 km^2 in the suburbs.

For each cell, we aggregate the number of CDRs at an hourly time scale to obtain a time series recording the level of activity on an hourly basis.

We normalize each ith cell's time series $x^i_{t=\text{day},h}$ with a z-score computed on an hourly basis. μ^i_h and σ^i_h are the 24 means and standard deviations of $x^i_{\text{day},h}$ for each hour. Thus, we obtain: $z^i_{\text{day},h} = (x^i_{\text{day},h} - \mu^i_h)/\sigma^i_h$. Using different μ^i_h and σ^i_h for different hours is very important because otherwise the circadian trend in our data would notably bias the synchro-

Figure 2 Example of daily rhythm in a mobile phone cell. (A) Original behaviour extracted from mobile phone data; (B) *z*-score scaled behaviour extracted from mobile phone data

nization among the time series (i.e., all time series would be highly synchronized because the day-night trend would cover more subtle differences).

The resulting time series (see Fig. 2) highlights deviations of the mean activity in different hours of the day on the one hand and on the other they are sufficiently stationary to apply standard statistics to measure the correlation (i.e., synchronization) of two time series.

For each NUTS-3 region, we compute two synchronization metrics: *within synchronization* is the average daily synchronization among cells assigned to the same municipality; *between synchronization* is the average daily synchronization among cells assigned to different municipalities (cells are assigned to municipalities based on the quantity of their overlapping area). Specifically, for each couple of cells i and j, we compute the average daily Mutual Information between $z_{\mathrm{day},h}^{i}$ and $z_{\mathrm{day},h}^{j}$: $\frac{1}{N} \sum_{\mathrm{day}=1}^{N} I(z_{\mathrm{day},h}^{i}; z_{\mathrm{day},h}^{j})$.

Mutual information is a natural measure of non-linear dependence quantifying the amount of information obtained about one time-series through the other one. Therefore, it measures how synchronized the two series are, and it is computed as:

$$I\bigl(z_{\mathrm{day},h}^{i}; z_{\mathrm{day},h}^{j}\bigr) = \int_{z_{\mathrm{day},h}^{i}} \int_{z_{\mathrm{day},h}^{j}} p\bigl(z_{\mathrm{day},h}^{i}, z_{\mathrm{day},h}^{j}\bigr) \log\left(\frac{p(z_{\mathrm{day},h}^{i}, z_{\mathrm{day},h}^{j})}{p(z_{\mathrm{day},h}^{i}) p(z_{\mathrm{day},h}^{j})} \right).$$

This approach computes a single average (*within* and *between*) synchronization for the whole time of observation (one month with our data). So, even if short-term events can spur sudden synchronization, the average value reflects longer-term trends in the behavioral patterns in the regions.

Figure 3 shows the distribution of *between* and *within* synchronization for the NUTS-3 regions under analysis. We consider the mean (among cells) of *between* and *within* synchronization as the reference value for each region (to be used in the regression model described below).

As aforementioned in the Introduction Section, we postulate that:

• High levels of *within synchronization* reflect the tendency of people to communicate together within their spatial cluster (i.e., municipality).

Figure 3 Violin plots, ordered by the median *within synchronization*, showing the average *between* and *within synchronization* of each city
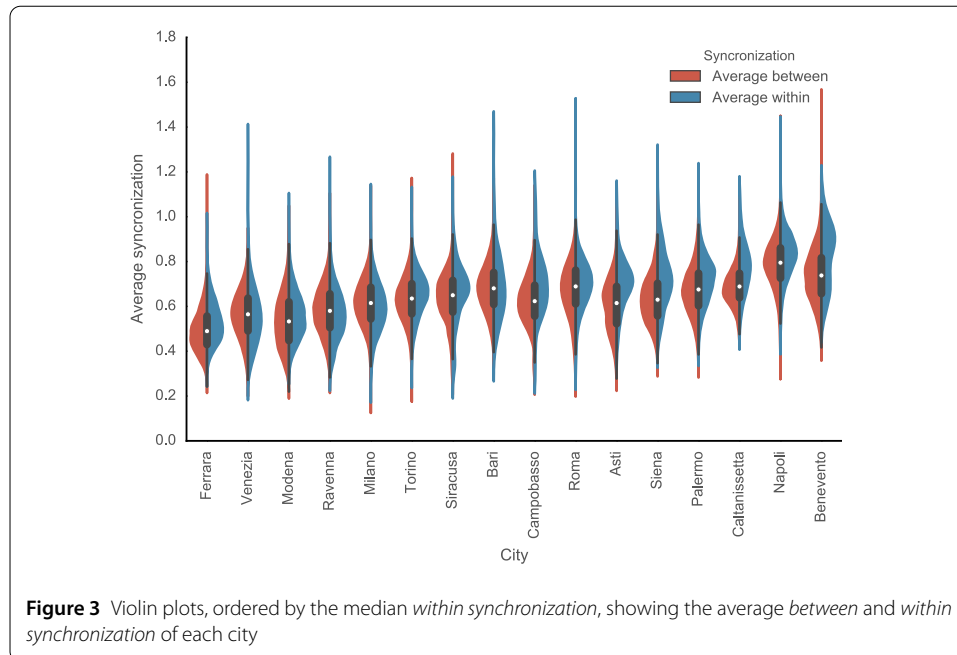
- High levels of *between synchronization* reflect instead the tendency of people to communicate together across different spatial clusters (i.e., municipalities).

We therefore use these two synchronization measures, computed from passively collected human behavioural data, to describe traditional proxies for social capital used in economics literature such as *Referendums turnout*, *Association density* and *Blood donations*.

In summary, for each of the 16 NUTS-3 regions under analysis, we compute the respective synchronization indices (i.e., *within* and *between* synchronization) and extract the traditional proxies for social capital. We check via Moran's I test that the obtained variables are not spatially auto-correlated, then we apply the linear regression analysis described in the following section.

2.1 Regression analysis

To validate our hypotheses, we describe the three social capital measures (i.e., Referendums turnout, Blood donations, and Association density) by means of three Ordinary Least Squares (OLS) models where the independent variables are: (i) *within synchronization*, (ii) *between synchronization*, and (iii) per-capita income. In principle many factors could affect the level of social capital and thus affect our estimation: the quality of institutions, the level of education, the degree of income inequality, to mention some. Following Alesina *et al.* [35] and Guiso *et al.* [36] we here consider per-capita income as a sole co-variate for the regression, to keep our estimates parsimonious, and use the level of per-capita income as a general proxy for these factors. Indeed higher per-capita income has been shown to be related to the strength of local institutions [37] and to the quality of education systems [18]. In Appendix C we report an additional set of regression analyses using the fraction of illiterate population, a good proxy for the level of education, as a sole covariate for the regression.

Between and *within synchronization* across NUTS-3 regions are highly correlated ($\rho = 0.9$), raising multicollinearity issues. Having correlated regressors, we have to rely on

multiple metrics to illustrate the statistical significance and importance of the variables in our model [38]. Thus, we report and discuss the variable importance through the beta weights, structure coefficients [39], commonality analysis components [40], dominance analysis [41] and Lindeman, Merenda, and Gold's (LMG) method [42].

Beta weights are often relied on to assess regressors' importance [39]. Beta weights indicate the expected increase/decrease in the dependent variable (e.g., Referendums turnout), expressed in standard deviation units, given a one standard deviation increase in such independent variable with all other independent variables held constant. However, the sole reliance on beta weights to interpret the contribution of each independent variable is justified only when the independent variables are perfectly uncorrelated [43]. In fact, beta weights may receive credit for explained variance shared with other regressors, while beta weights of the other regressors are not given credit for this shared variance [43]. Therefore, the contribution of the other regressors to the regression effect may be not fully captured. Moreover, beta weights have also limitations in determining *suppression* effects in a regression, that is, a regressor that contributes little or no variance to the dependent variable but it may have a large non-zero beta weight because it *purifies* one or more regressors of their irrelevant variance, thereby increasing its or theirs predictive power [44].

Structure coefficients quantify the strength of the bi-variate relationship between each regressor and the dependent variable in isolation from other correlations between regressors and dependent variable. Hence, they are a useful measure of the *direct* effect of a regressor [39]. Being only a measure of direct effect, they are unable to identify regressors sharing explained variance in the dependent variable, and thus to quantify the amount of this shared variance [39]. Instead, the LMG measure can be thought as the average improvement of regressor X_1, over all models of size s without X_1 [42].

In order to quantify the contribution that each regressor shares with every other set of regressors, we also perform a commonality analysis [40]. This technique decomposes R^2, and thus the total effect (Tot_{CA}), into its *unique* (U_{CA}) and *common* (C_{CA}) effects. Unique effects indicate how much variance is uniquely accounted for by a single regressor; while common effects indicate how much variance is common to each set of regressors [40]. It is worth noting that if the regressors are all uncorrelated, the contributions of all regressors are unique effects, as no variance is shared between independent variables in the prediction of the dependent variable.
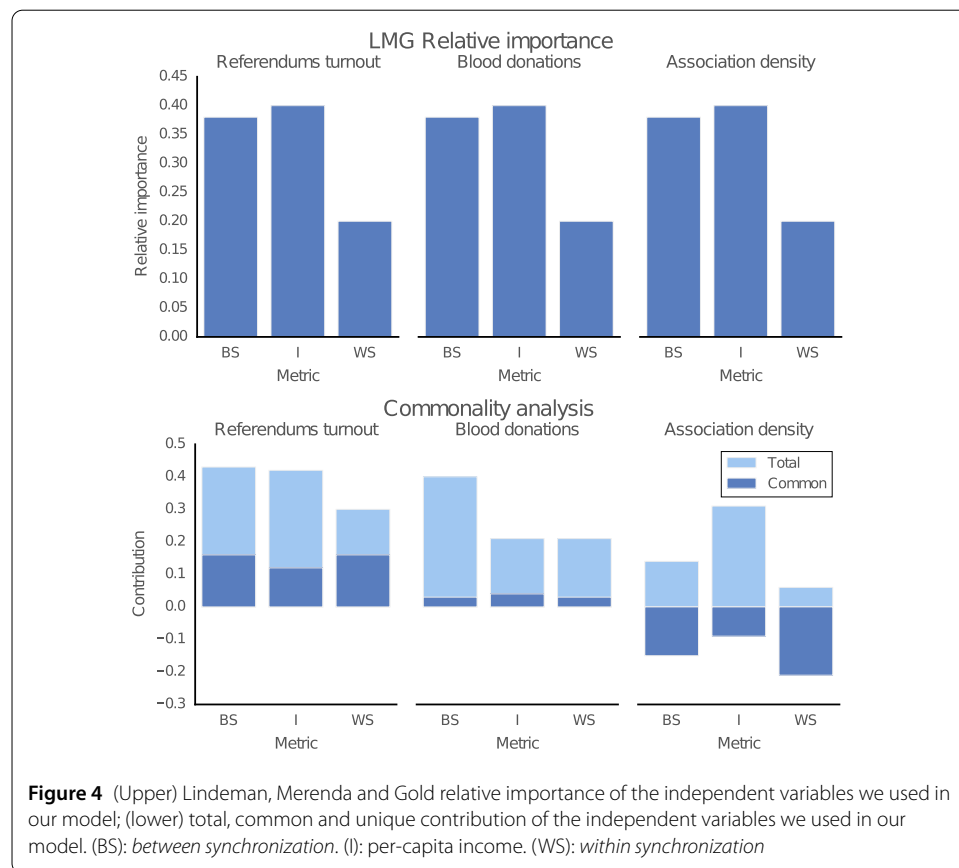
Moreover, we use dominance analysis [41] to determine the importance of a regressor based on comparisons of unique variance contributions of all pair of independent variables to regression equations involving all possible subsets of regressors. Interestingly, dominance analysis is a technique able to quantify (i) the *direct* effect of a regressor in isolation from other regressors, as the subset containing no other regressors includes zero-squared correlations, (ii) the *total* effect, as it compares the unique variance contributions of the regressors when all of them are included in the model, and (iii) the *partial* effect, as it compares the unique variance contributions of the regressors for all the possible subsets of them.

3 Results

Results of OLS models are shown in Table 1, where we report the adjusted R^2_{adj}[b] of the OLS using *between* synchronization, *within* synchronization and per-capita income as covariates.

Table 1 Referendums turnout, Blood donations, Association density represented by *between* and *within synchronization*, controlled for per-capita income were tested using commonality analysis. As for statistical significance of the beta weights, we use the following notation: *$p < 0.05$, **$p < 0.01$

	β (95% CI)	r_s	U_{CA}	C_{CA}	Tot_{CA}	LMG
Referendums turnout (R^2_{adj}: 0.68)						
Between sync	−0.12** (−0.20, −0.05)	−0.76	0.27	0.16	0.43	0.38
Within sync	0.09* (0.01, 0.18)	−0.63	0.13	0.16	0.30	0.20
Per-capita income	0.06** (0.02, 0.10)	0.75	0.30	0.12	0.42	0.40
Blood donations (R^2_{adj}: 0.55)						
Between sync	−24.91** (−40.44, −9.37)	−0.79	0.36	0.03	0.40	0.52
Within sync	19.49* (2.45, 36.54)	−0.58	0.18	0.03	0.21	0.24
Per-capita income	8.49* (0.67, 16.31)	0.57	0.16	0.04	0.21	0.22
Association density (R^2_{adj}: 0.52)						
Between sync	−21.88** (−37.54, −6.23)	−0.48	0.29	−0.15	0.14	0.30
Within sync	22.96* (5.78, 40.14)	−0.31	0.27	−0.21	0.06	0.27
Per-capita income	13.00** (5.12, 20.88)	0.71	0.41	−0.09	0.31	0.42

Figure 4 (Upper) Lindeman, Merenda and Gold relative importance of the independent variables we used in our model; (lower) total, common and unique contribution of the independent variables we used in our model. (BS): *between synchronization*. (I): per-capita income. (WS): *within synchronization*

The variable importance of the independent variables is reported through the Beta weights, the structure coefficients [39], the commonality analysis components [40], the dominance analysis [41] and the Lindeman, Merenda, and Gold's (LMG) method [42]. Figure 4 summarizes the results of two of the most used variable importance metrics.

Table 2 Referendums turnout: Dominance analysis output. The ✓ symbol represents the dominance of a variable *A* on *B*. The × symbol represents the dominance of a variable *B* on *A*. In empty cells dominance could not be established between regressors

Dominance	Complete	Conditional	General
Between sync > within sync	✓	✓	✓
Between sync > per-capita income			×
Within sync > per-capita income	×	×	×

Table 3 Blood donations: Dominance analysis output. The ✓ symbol represents the dominance of a variable *A* on *B*

Dominance	Complete	Conditional	General
Between sync > within sync	✓	✓	✓
Between sync > per-capita income	✓	✓	✓
Within sync > per-capita income	✓	✓	✓

Here we provide a detailed analysis of each social capital proxy used in economics literature.

Referendums turnouts. The first group of rows of Table 1 shows that *between synchronization* contributes the most to the regression equation ($\beta = -0.12$), while holding all other regressors constant. It is the most correlated variable with the predicted *Referendums turnout* ($r_s = -0.76$) and the major contributor to the regression effect ($Tot_{CA} = 0.43$), where 27.2% of regression effects is *unique* and 16.2% is in common with the other variables. The relative importance of *between synchronization* ($Tot_{CA} = 0.43$ and LMG = 0.38) is closely related to the one of *per-capita income* ($Tot_{CA} = 0.42$ and LMG = 0.40). Dominance analysis confirms this importance (see Table 2).

The second most important beta weight is *within synchronization* that, besides its positive value, has negative correlation with *Referendums turnout* ($r_s = -0.63$). This may indicate that the regression effect was confounded by all the variables included in the model but they all contribute substantially in the explanation of *Referendums turnout* (all C_{CA} and Tot_{CA} values are greater than zero).
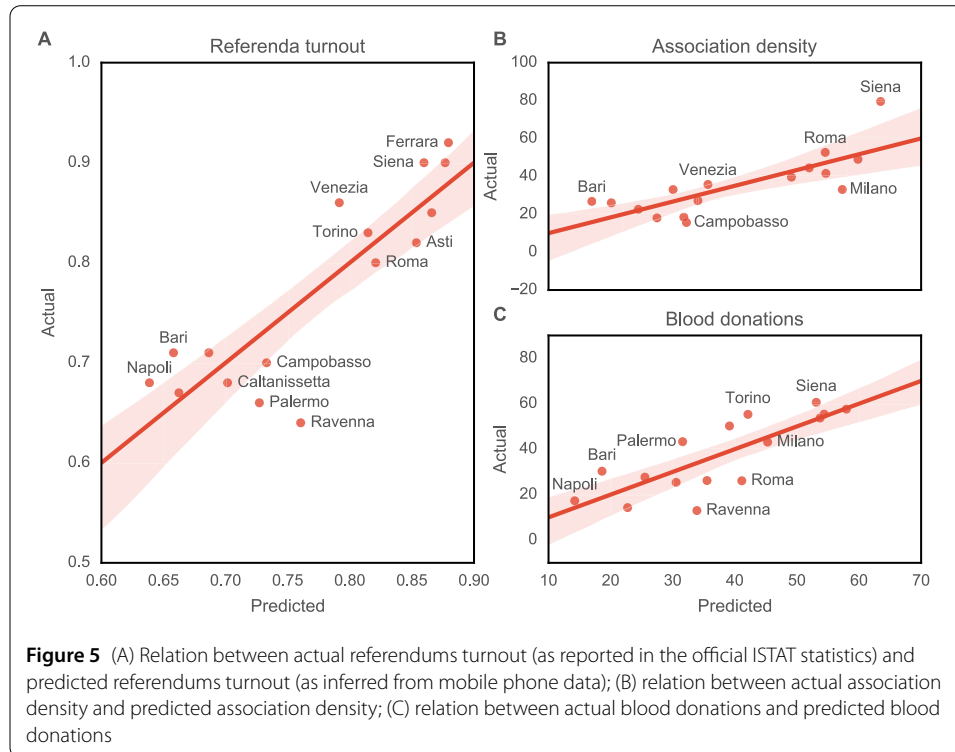
Blood donations. From the second group of rows of Table 1 we observe that *between synchronization* holds the highest contribution to the regression in all the metrics, accounting for 52% of the importance in the model ($\beta = -24.91$), highest *total* ($Tot_{CA} = 0.40$) and *unique* contribution ($U_{CA} = 0.36$).

The second most important beta weight is *within synchronization* that, besides its positive value, has negative correlation with *Blood donations* ($r_s = -0.580$). This may indicate that the regression effect was confounded by all the variables included in the model but they all contribute substantially in the explanation of *Blood donations* (all C_{CA} and Tot_{CA} values are greater than zero). The importance of *within synchronization* is very close to the importance of *per-capita income*, but from the Dominance analysis (see Table 3) we have that *per-capita income* has a minor role in the regression.

Associations density. The last group of rows in Table 1 shows that *within synchronization* and *between synchronization* obtained the largest beta weights ($\beta = 22.96$ and $\beta = -21.88$ respectively), demonstrating the most important contributions to the regression equation, while holding all other regressors constant. Despite this, *per-capita income* accounts for 42% of the importance in the model, having also the highest *total* ($Tot_{CA} = 0.42$) and *unique* contribution ($U_{CA} = 0.41$). From the Dominance analysis (see Table 4) it is possible

Table 4 Association density: Dominance analysis output. The ✓ symbol represents the dominance of a variable *A* on *B*. The × symbol represents the dominance of a variable *B* on *A*

Dominance	Complete	Conditional	General
Between sync > within sync	✓	✓	✓
Between sync > per-capita income	×	×	×
Within sync > per-capita income	×	×	×

Figure 5 (A) Relation between actual referendums turnout (as reported in the official ISTAT statistics) and predicted referendums turnout (as inferred from mobile phone data); (B) relation between actual association density and predicted association density; (C) relation between actual blood donations and predicted blood donations

to see that the most important variable is indeed *per-capita income*, followed by *between synchronization* and *within synchronization*.

Particularly, besides the positive value of *within synchronization*'s beta weight, it is negatively correlated with *Association density* ($r_s = -0.31$). Together, the very small structure coefficient ($r_s^2 = 0.09$) and the negative common effect ($C_{CA} = -0.21$) may indicate [45] the suppression role of *within synchronization* in the regression that *purifies* the variance explained by the other variables.

4 Discussion

Taken together, our results show that the models can explain the 68% of the variation in *Referendums turnout* ($R_{adj}^2 = 0.68$), the 55% of the variation in *Blood donations* ($R_{adj}^2 = 0.55$) and the 52% of the variation in *Association density* ($R_{adj}^2 = 0.52$). Figure 5 shows the distribution of the fitted points.

Particularly, *within synchronization* correlates positively with social capital metrics ($\beta = 0.09$ for Referendums turnout, $\beta = 19.49$ for Blood donations, and $\beta = 22.96$ for Association density). Thus, this indicator informs us on the intensity of cohesion within close-proximity groups and communities, which approximates *"...the instantiated informal norm that promotes co-operation between two or more individuals...* [18]*"*.

In Larssen et al., individuals with strong social bonding (i.e., association and trust among neighbors) are more likely to take civic action.
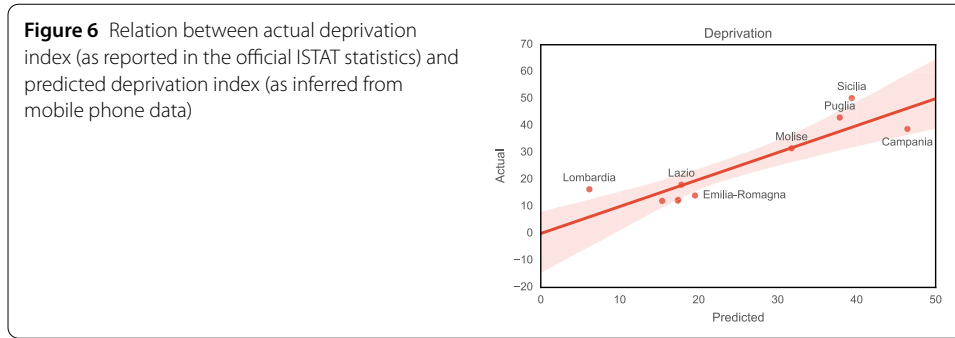
Our second indicator, *between synchronization*, captures the tendency of a given community (i.e., a given municipality) to communicate with many different communities (i.e., other municipalities). Thus, more *between synchronization* implies more interaction among multiple groups (i.e., municipalities); while less *between synchronization* implies less interaction and more isolation among groups. Interestingly, our results correlate negatively a high level of *between synchronization* with standard social capital metrics ($\beta = -0.12$ for Referendums turnout, $\beta = -24.91$ for Blood donations, and $\beta = -21.88$ for Association density). These findings are in line with a number of theoretical and empirical works claiming that diversity undermines a sense of community and social cohesion [20, 35, 46–49]. For example, Alesina and La Ferrara [46] have studied whether and how much the degree of heterogeneity in communities influences the amount of participation in different types of groups. Using survey data on group membership and data on localities in United States, they found that, after controlling for many individual characteristics, participation in associations (e.g., religious groups, hobby clubs, youth and sport groups, etc.) is significantly lower in more different, unequal, and racially or ethnically fragmented localities.

Our results are obtained including per-capita income in the regressions, similarly to what is done in the literature [22, 35]; controlling for wealth at the level of the NUTS-3 regions. The role of per-capita income is indeed important. We find that per-capita income has a strong relevance in describing the Association density, while it shows a minor role in explaining the higher Referendums turnout and Blood donations.

5 Conclusion

In this paper, we have introduced a couple of novel synchronization metrics (i.e., *within* and *between synchronization*) that represent an innovative and efficient way to describe traditional social capital measures (i.e., Referendum turnouts, Blood donations, and Association density). The proposed approach is, at the best of our knowledge, the first one that combines synchronization metrics and mobile phone data, which are always up to date and available for a very large fraction of the world population. A further merit of our approach is the ability to identify and analyze individually the role played by the level of cooperation within a close proximity-based community (i.e., *within synchronization*), and the one played by the level of cooperation among different communities in a larger geographical area (i.e., *between synchronization*). Moreover, our approach does not need individual-level data, which is rarely shared by telecommunication operators to ensure data confidentiality. It is also worth noting that our synchronization-based approach can be extended easily to other sources of information such as activities on social media platforms, mobility routines captured from transportation data, etc.

Social capital is a key determinant to understand neighborhood stability for crime prevention, to enforce social cohesion, e.g., immigrant integration, and to create integration tools ind addition to language and culture training. Thus, the geographical characterization of areas with differential levels of social capital is an important tool in the hands of policy makers aiming at specific incentive policies, which are clearly more or less effective depending on the underlying social capital types and levels.

Figure 6 Relation between actual deprivation index (as reported in the official ISTAT statistics) and predicted deprivation index (as inferred from mobile phone data)

Table 5 Deprivation represented by *between* and *within synchronization*, controlled for per-capita income was tested using commonality analysis

	β (95% CI)	r_s	r_s^2	U_{CA}	C_{CA}	Tot_{CA}	LMG
R_{adj}^2: 0.68							
Between sync	10.84 (−4.36, 26.05)	0.76	0.58	0.10	0.35	0.46	0.27
Within sync	−4.92 (−21.68, 11.83)	0.76	0.58	0.01	0.44	0.46	0.19
Per-capita income	−9.91 (−18.10, −1.75)	−0.85	0.73	0.30	0.27	0.57	0.52

Table 6 Deprivation Dominance analysis output. The ✓ symbol represents the dominance of a variable A on B. The × symbol represents the dominance of a variable B on A

| Dominance	Complete	Conditional	General
Between sync > within sync	✓	✓	✓
Between sync > per-capita income	×	×	×
Within sync > per-capita income	×	×	×

Appendix A: Regression using Multiple Deprivation Index

While there seems to be a growing empirical evidence that social capital contributes significantly to sustainable development, a number of authors raise issues and point to unconvincing and conflicting results [30, 50]. At the heart of the problem is the multiple definitions and metrics of both social capital and sustainable development. Following this line of research and other similar works [13, 14], we analyze the association between our synchronization metrics (i.e., *within* and *between synchronization*) and the Multiple Deprivation Index, see Fig. 6. Multiple Deprivation Index is a synthetic measure used for analyzing social exclusion. It combines information comprising household structure, level of education and participation in the labour market. Our data is based on official ISTAT statistics and refer to year 2013.

Having the deprivation data available only at the NUTS-2 region level, the regression is applied only to few data-points. This issue causes high instability of the coefficients of the OLS regression (see 95% CI column of Table 5). For this reason we show here the results of the analysis (Table 5) and the dominance results (Table 6) without deep explanations. Nevertheless, the explained variance is very high, meaning that this associative relation should be further investigated in future studies.

Appendix B: Correlation matrix among variables

To present the described correlation and dominance analysis in a more intuitive way, in Fig. 7, we report the correlation matrix among all the variables. It is possible to see that R^2

Figure 7 Correlation matrix between all the variables. The upper panel reports R^2 among pairs
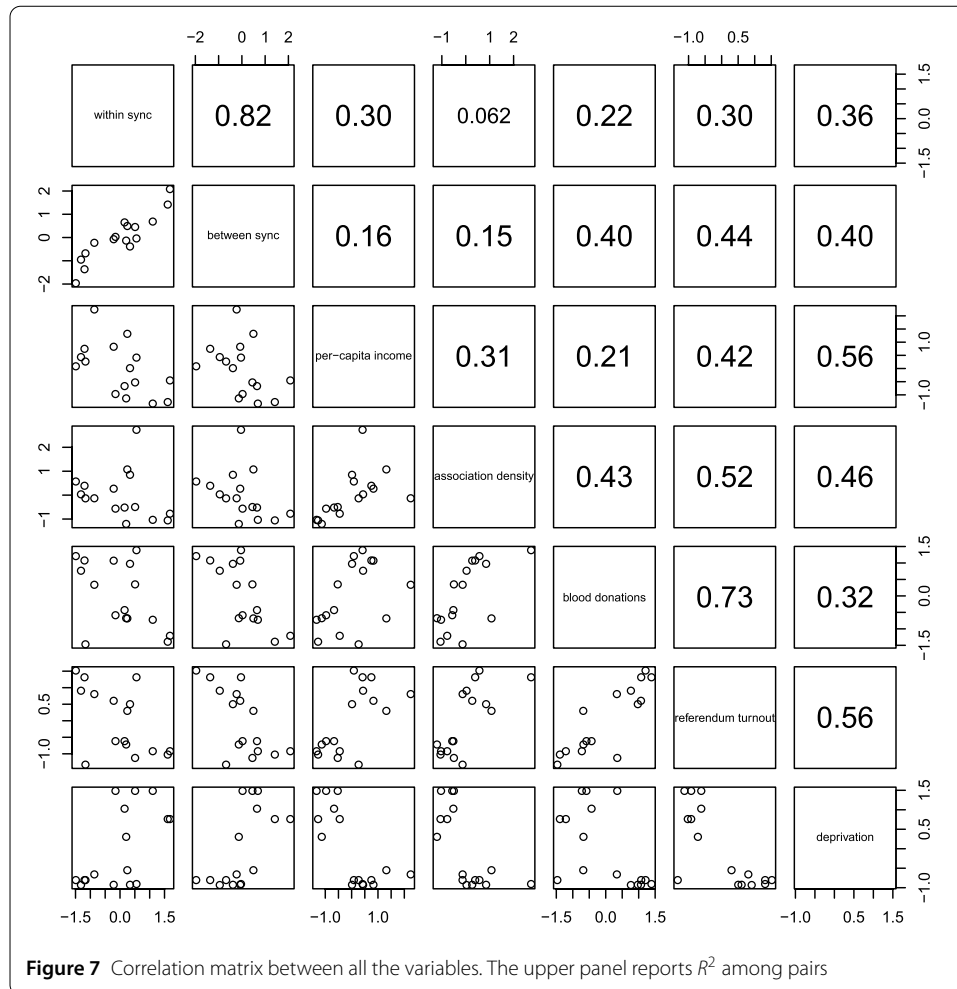
among pairs is lower than in the multiple regression case. The per-capita income has an important role as confounding factor (and has been included into the covariates for this reason), but by no means it is able to explain the regression alone.

Appendix C: Testing the robustness

We conduct some additional analyses to test the robustness of our approach. Firstly, we verify the impact of the temporal aggregation used to compute the (*within* and *between*) synchronization values. While in the main text, we use CDR counts aggregated using a 1-hour temporal window, in Table 7 we also report the results obtained with a 2-hours temporal window. Results remain similar, but due to the limited number of data points we often lose statistical significance.

A set of additional regression analyses tests whether a different covariate, i.e., the fraction of illiterate population, can substitute per-capita income in the regression model. Table 8 shows the obtained results. It is interesting to see that the use of this covariate does not change the basic structure of our regression: positive correlation with *within synchronization*, negative correlation with *between synchronization*, although statistical significance is weaker than in the case of per-capita income. This can be partially explained by the low number of data-points, which can influence the *p*-value.

Table 7 Regression results obtained with a 2-hours time window. As for statistical significance of the beta weights, we use the following notation: *$p < 0.05$, **$p < 0.01$

	Referendum turnout	Blood donations	Association density
Between sync	−0.08*	−18.97*	−16.68*
Within sync	0.05	13.21	19.11*
Per-capita income	0.06*	7.78	13.78**
R^2_{adj}	0.54	0.43	0.42

Table 8 Referendums turnout, Blood donations, Association density represented by *between* and *within synchronization*, controlled for population illiteracy. As for statistical significance of the beta weights, we use the following notation: *$p < 0.05$, **$p < 0.01$

	Referendum turnout	Blood donations	Association density
Between sync	−0.09*	−21.15*	−16.44*
Within sync	0.08	17.51	23.26**
Pop. illiterate	−0.06**	−8.46	−17.26**
R^2_{adj}	0.61	0.50	0.69

List of abbreviations
CDRs, Call Detail Records; NUTS-3, Nomenclature des unités territoriales statistiques, level 3; LAU-2, Local Administrative Units, level 2; OLS, Ordinary Least Squares; Tot_{CA}, total effect; U_{CA}, unique effects; C_{CA}, common effects; LMG, Lindeman, Merenda, and Gold's; CI, Confidence Intervals; r_s, structure coefficient; within-sync, *within synchronization*; between-sync, *between synchronization*.

Availability of data and materials
The majority of the data sources supporting the conclusions of this article are included within the article (and its additional file(s)). The mobile phone data provided by TIM were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of TIM. We analyzed the data thanks to freely available R packages: i) standard R linear model (`lm`) for OLS; ii) Package `spdep` for spatial auto-correlation analysis via Moran's I test; iii) Packages `boot`, `yhat`, `realimpo` to analyze the contribution of multiple regressors. In particular, `yhat` provides methods to interpret multiple linear regression and canonical correlation results. `relaimpo` provides several metrics for assessing relative importance in linear models. The source code to repeat the experiments of this article is available at `https://github.com/mmamei/socialk`.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
Conceived the study: MM, FP. Designed and performed the experiments: MM, FP, MDN, BL. Analyzed the data: MM, FP, MDN, BL. Wrote the paper: MM, FP, MDN, BL. All authors read, reviewed and approved the final manuscript.

Author details
[1] University of Modena and Reggio Emilia, Modena, Italy. [2] University of Trento, Povo, Italy. [3] Fondazione Bruno Kessler, Povo, Italy. [4] SKIL-TIM, Povo, Italy. [5] Massachusetts Institute of Technology, Cambridge, USA.

Endnotes
[a] For a given phone call or SMS exchange we record only the CDR from the originating mobile terminal.
[b] The adjusted R^2_{adj} is a variant of the R^2 that aims at overcoming the spurious increase of the former when extra variables are added to the model. It is defined as $R^2_{\text{adj}} = 1 - (1 - R^2)\frac{n-1}{n-k-1}$ where n is the number of data-points and k the number of parameters in the model.

Publisher's Note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References
1. Sumpter DJ (2006) The principles of collective animal behaviour. Philos Trans R Soc Lond B, Biol Sci 361(1465):5–22

2. Schneidman E, Berry MJ, Segev R, Bialek W (2006) Weak pairwise correlations imply strongly correlated network states in a neural population. Nature 440:1007–1012
3. Strogatz SH (2003) Sync: the emerging science of spontaneous order. Theia, New York
4. Neda Z, Ravasz E, Brechet Y, Vicsek T, Barabasi A-L (2000) Self-organizing processes: the sound of many hands clapping. Nature 403:849–850
5. Saavedra S, Hagerty K, Uzzi B (2010) Synchronicity, instant messaging, and performance among financial traders. Proc Natl Acad Sci USA 108(13):5296–5301
6. Hong L, Page SE (2004) Groups of diverse problem solvers can outperform groups of high-ability problem solvers. Proc Natl Acad Sci USA 101(46):16385–16389
7. Nowak MA (2006) Five rules for the evolution of cooperation. Science 314(5805):1560–1563. https://doi.org/10.1126/science.1133755. http://science.sciencemag.org/content/314/5805/1560
8. Wiltermuth SS, Heath C (2009) Synchrony and cooperation. Psychol Sci 20(1):1–5
9. McNeill WH (1997) Keeping together in time. Harvard University Press, Cambridge
10. Hannah JL (1977) African dance and the warrior tradition. J Asian Afr Stud 12(1–4):111–133
11. Ehrenreich B (2007) Dancing in the streets: a history of collective joy. Metropolitan Books, New York
12. Schläpfer M, Bettencourt L, Grauwin S, Raschke M, Claxton R, Smoreda Z, West G, Ratti C (2014) The scaling of human interactions with city size. J R Soc Interface 11:20130789
13. Eagle N, Macy M, Claxton R (2010) Network diversity and economic development. Science 328(5981):1029–1031. https://doi.org/10.1126/science.1186605. http://science.sciencemag.org/content/328/5981/1029
14. Blumenstock J, Cadamuro G, On R (2015) Predicting poverty and wealth from mobile phone metadata. Science 350(6264):1073–1076
15. De Nadai M, Staiano J, Larcher R, Sebe N, Quercia D, Lepri B (2016) The death and life of great Italian cities: a mobile phone data perspective. In: Proceedings of the 25th international conference on world wide web, pp 413–423
16. Wesolowski A, Eagle N, Tatem AJ, Smith DL, Noor AM, Snow RW, Buckee CO (2012) Quantifying the impact of human mobility on malaria. Science 338(6104):267–270
17. Bogomolov A, Lepri B, Staiano J, Letouzé E, Oliver N, Pianesi F, Pentland A (2015) Moves on the street: classifying crime hotspots using aggregated anonymized data on people dynamics. Big Data 3(3):148–158
18. Fukuyama F (2001) Social capital, civil society and development. Third World Q 22(1):7–20
19. Putnam RD, Leonardi R, Nanetti R (1993) Making democracy work: civic traditions in modern Italy. Princeton University Press, Princeton
20. Knack S, Keefer P (1997) Does social capital have an economic payoff? A cross-country investigation. Q J Econ 112(4):1251–1288. https://doi.org/10.1162/003355300555475
21. Fukuyama F (1995) Trust: the social virtues and the creation of prosperity. The Free Press, New York
22. Guiso L, Sapienza P, Zingales L (2004) The role of social capital in financial development. Am Econ Rev 94(3):26–556
23. Banfield EC, Fasano L (1958) The moral basis of a backward society. The Free Press, New York
24. Nannicini T, Stella A, Tabellini G, Troiano U (2013) Social capital and political accountability. Am Econ J Econ Policy 5(2):222–250
25. Putnam RD (2000) Bowling alone: the collapse and revival of American community. A Touchstone book. Simon & Schuster, New York
26. Woolcock M, Narayan D (2000) Social capital: implications for development theory, research, and policy. World Bank Res Obs 15(2):225–249
27. Norbutas L, Corten R (2018) Network structure and economic prosperity in municipalities: a large-scale test of social capital theory using social media data. Soc Netw 52(1):120–134
28. Whiteley PF (2000) Economic growth and social capital. Polit Stud 48(3):443–466
29. Paxton P (1999) Is social capital declining in the United States? A multiple indicator assessment. Am J Sociol 105(2):88–127
30. Bigoni M, Bortolotti S, Casari M, Gambetta D, Pancotto F (2016) Amoral familism, social capital, or trust? The behavioural foundations of the Italian North–South divide. Econ J 126:1318–1341. https://doi.org/10.1111/ecoj.12292
31. Guiso L, Sapienza P, Zingales L (2010) Civic capital as the missing link. In: Handbook of social economics, vol. 1, pp 417–480
32. Guiso L, Sapienza P, Zingales L (2009) Cultural biases in economic exchange? Q J Econ 124(3):1095–1131
33. Buonanno P, Montolio D, Vanin P (2009) Does social capital reduce crime? J Law Econ 52(1):145–170
34. Cartocci R (2007) Mappe del tesoro: atlante del capitale sociale in Italia. Il mulino, Bologna
35. Alesina A, La Ferrara E (2002) Who trusts others? J Public Econ 85(2):207–234
36. Guiso L, Sapienza P, Zingales L (2016) Long-term persistence. J Eur Econ Assoc 14(6):1401–1436
37. Helliwell JF, Putnam RD (1995) Economic growth and social capital in Italy. East Econ J 21(3):295–307
38. Nathans LL, Oswald FL, Nimon K (2012) Interpreting multiple linear regression: a guidebook of variable importance. Pract Assess Res Eval 17:9
39. Courville T, Thompson B (2001) Use of structure coefficients in published multiple regression articles: β is not enough. Educ Psychol Meas 61(2):229–248
40. Rowell RK (1991) Partitioning predicted variance into constituent parts: how to conduct commonality analysis
41. Azen R, Budescu DV (2003) The dominance analysis approach for comparing predictors in multiple regression. Psychol Methods 8(2):129
42. Lindeman R (1980) Introduction to bivariate and multivariate analysis. Scott, Foresman and Company, Glenview
43. Pedhazur EJ (1997) Multiple regression in behavioral research: explanation and prediction. Harcourt Brace, New York
44. Capraro RM, Capraro MM (2001) Commonality analysis: understanding variance contributions to overall canonical correlation effects of attitude toward mathematics on geometry achievement. Mult Linear Regres Viewp 27:16–23
45. Kerlinger FN, Pedhazur EJ (1973) Multiple regression in behavioral research. Holt, Rinehart and Winston, New York
46. Alesina A, La Ferrara E (2000) Participation in heterogeneous communities. Q J Econ 115(3):847–904
47. Glaeser E, Laibson D, Scheinkman J, Soutter C (2000) Measuring trust. Q J Econ 115(3):811–846
48. Costa DL, Kahn ME (2002) Civic engagement and community heterogeneity: an economist's perspective. Perspective Polit 1(1):103–111

49. Miguel E, Gugerty MK (2005) Ethnic diversity, social sanctions, and public goods in Kenya. J Public Econ 89(11–12):2325–2368
50. Sabatini F (2008) Social capital and the quality of economic development. Kyklos 61(3):466–499