The underlying geometry of organizational dynamics: similarity-based social space and labor flow network communities

Hernan Mondani¹

Published online: 13 November 2017

© The Author(s) 2017. This article is an open access publication

Abstract In this article, we use Swedish longitudinal register data to study the effect that similarity in organizational properties has on the interaction between organizations. We map out the social space of large organizations in the Stockholm Region and the interplay between social distance and the network communities of employee movements between organizations. We firstly use homogeneity analysis to describe the dynamics of organizations in terms of the time evolution of their similarity. Our results show that most categorical variables are quite stable over time. Organizations linked through employee movement edges have a lower average distance in social space than non-linked organizations. Secondly, we look at network community dynamics in social space. Employee flows between organizations in different communities exhibit a so-called gravity law from spatial statistics, decaying more slowly than observed geographical networks, meaning that employees reach out regions of social space further than of physical space. Finally, the rate of change of distance in homogeneity space exhibits a statistical distribution similar to the ones found in various other growth processes in natural and man-made systems.

Keywords Organizational dynamics · Similarity · Social space · Social distance · Homogeneity analysis · Network community structure · Growth process

✉ Hernan Mondani
hernan.mondani@sociology.su.se

¹ Department of Sociology, Stockholm University, Universitetsvägen 10B, 106 91 Stockholm, Sweden

1 Introduction

1.1 Social space and the geometry of interactions

Social life unfolds in physical space. The complex interactions between individuals, institutions and organizations are allowed as well as constrained by space. It is known, for example, that social interaction patterns like friendship networks are embedded in the geographical setting of the people involved, in such a way that the likelihood of a contact tie to form between two individuals decreases as the geographical distance separating them increases (Butts et al. 2012). Likewise, a well-studied fact in urban systems is that the amount of people traveling between two cities is roughly inversely proportional to the distance separating the cities (Zipf 1946).

But society is not bounded by physical space alone. Social interaction over time gives rise to an array of specific constraints not necessarily captured by physical constraints. Status hierarchies are a good example of an object with structure-like properties that are—at least not entirely—of physical origin. An office clerk may work right next to the company director, yet a series of nonphysical social barriers and conventions probably affect the likelihood to meet and become friends. Physical space might be the substrate, but social space plays a role in the likelihood.

It is within this context of interplay between physical and social properties that the notion of *social space* becomes useful. Proximity in social space is related to some form of *similarity* along a set of relevant social characteristics. The position of two individuals in social space allows defining a *social distance* by analogy with geographical distance. The assumption is that proximity of two individuals in social space increases the likelihood of them interacting, and conversely, their interaction decays as they are placed further apart in social space. This assumption is grounded on one of the most robust empirical findings in social network analysis: the fact that similar individuals tend to be linked among themselves to a higher degree than dissimilar individuals, something generally known as *homophily* (McPherson et al. 2001). Through the notion of social distance, a geometrical view on social interaction is introduced.

Researchers have advanced different notions of social space. Blau (1977) conceptualized it on the basis of sociodemographic properties, e.g. age, sex, education level, socioeconomic status, etc. Each of these individual-level attributes becomes a dimension in a multidimensional social space, essentially an axis in a Cartesian coordinate system. Blau saw this space as an adequate representation social structure at the societal level, imposing constraints on the types of interaction patterns that are likely to occur in a population. Due to this seminal contribution McPherson (2004) coined the term *Blau space* to refer to such type of social spaces.

Blau's conceptualization of social space defines a priori theoretically derived dimensions of the space, with as many coordinate axes as relevant sociodemographic variables. Another important and qualitatively very different conceptualization is not to assume any structure a priori, and let the geometry be determined by

the similarity between category values. This approach, relational in essence, is probably best known among sociologists because of ‘*Distinction: a social critique of the judgment of taste*’ (Bourdieu 1984). He used similarity—called correspondence—between social variables in order to map out the structure of a social field. Within this framework, individuals in different regions of this space possess different kinds of capital, and interaction between individuals and groups of individuals such as classes unfolds along the constraints of the space (Bourdieu 1985, 1987).

Before moving forward, we should mention a critical issue in this kind of similarity-based dynamics: the difficulty in teasing out whether similarity precedes interaction or vice versa. That is to say, do two social entities (individuals, organizations) begin to interact because they are similar—i.e. close in social space—or do interacting social entities become closer more similar—and consequently closer in social space—by virtue of their interaction? We will come back to this point and check it with our data.

1.2 Organizational dynamics in social space: similarity and communities

The demographic perspective on organizational dynamics has the organization as the unit of analysis. Researchers have studied the birth and death of organizations and its impact on a whole population of organizations (Carroll and Hannan 2000). Within these organizational dynamical processes, similarity plays a role as well. Organizations either group themselves or become grouped into sectors and industries conformed by organizations with similar characteristics, people tend to change jobs between organizations in similar sectors and industries.

The previous point regards organizational dynamics at the level of organizations. At a more micro level of the individual, it has long been known that organizations to a certain extent hire employees through other employee’s recommendations, and that job seekers on the other hand make use of contacts in the proximity of their social networks to apply for positions, making networks an important factor for employee mobility in the labor market (Montgomery 1991).

It is therefore not surprising that ideas and models based on some form of social space have been applied to organizational dynamics questions. McPherson and Ranger-Moore (1991) focus on an application to voluntary organizations, introducing a social space model à la Blau where education and occupational status (or prestige) are the two central Blau dimensions, and they together with the available organizational capacity create an organizational landscape that evolves over time. Recruitment for voluntary organizations takes place in each person’s neighborhood in Blau space.

Closely related to social space and organizations is the idea that social space neighborhoods can offer a proxy for an organization’s *social niche*. The concept of niche has been applied to the study of social movements (Stern 1999) and political landscapes of a social group in a geographical area as shown in Leuthold et al. (2007). Such studies, although incorporating a dynamical viewpoint, have lacked a network perspective. Additionally, they tend to be à la Blau and thus rely on an assume a priori a geometry defined by preset dimensions.

On the other hand, there are two approaches to analyze how organizational populations interact: population ecology and community ecology. The first approach is mainly due to Hannan and Freeman (1977) and later Hannan and Freeman (1989). The second approach is advanced by Astley (1985) in a critique of Hannan and Freeman's original formulation of organizational ecology. An *organizational community* is then an environment where different populations of organizations coexists and engage in different interaction processes such as competition and symbiosis (Aldrich and Ruef 2006, Chap. 11). Of interest is the stability of communities over time, and the differentials between the dynamics within communities and the dynamics between communities. In this study, we explore community stability by means of the flow of employees between organizations.

There might be different aspects to social space that are not captured by a single logic. That is to say, there could be more than one aspect to social space, and different aspects could coexist and interplay. In this study, we explore a combination of two aspects of the social space of organizations: the demographic aspect, which we study by homogeneity analysis without assuming any orthogonality in the relevant dimensions of social space, and the community aspect, which we incorporate by using network community detection techniques to estimate groups of labor flows more tightly connected among each other.

1.3 Aim of this study

In this article, we study the effect of similarity in organizational properties on the interaction between organizations. Concretely, we investigate to what extent more similar organizations are more likely to be linked to each other, and whether belonging to the same organizational community makes a difference in the social space of interaction.

To that end, we use Swedish longitudinal register data and map out a similarity-based social space of large organizations in the Stockholm Region and the interplay between social distances and the network communities of employee inter-organizational movements. The data is quite unique, since it allows for the tracking over time of all employees and organizations, and through that the construction of an inter-organizational network. On the methodological side, we use a combination of homogeneity analysis (a similarity-based dimensionality reduction) and network community structure algorithms in order to provide a geometrical interpretation to organizational dynamics and address different aspects of the description of social space.

The remainder of this article is structured as follows. We describe the database in Sect. 2. The methods and visual tools we use are outlined in Sect. 3. Section 4 presents the analyses. Finally, in Sect. 5 we provide some concluding remarks and discuss lines of further research.

2 Data

Our study is not data-driven, but rather intends to apply a new combination of methods—homogeneity analysis and network community detection—to a suitable data set. Thus, we use the Stockholm database for our analyses. It is a compilation of Swedish governmental registers compiled by Statistics Sweden (*Statistiska centralbyrån*, SCB).¹ It contains information on organizations in the Stockholm Region for the period 1990–2003.

The period of analysis is chosen due to the availability of data, but it is at the same time an interesting period to study, for historical reasons. It is characterized by important macroeconomic transformations like the Swedish economic crisis at the beginning of the 1990s, the shrinking of the public sector and the shift from manufacturing to more service-oriented private activities (Bergmark and Palme 2003). The data also provide an opportunity to try this innovative combination of methods.

All organizations with employees registered in the Region are included. The legal ownership allows distinguishing between public and private organizations, and we also have information on their main activity at the workplace level. Each organization has a unique, de-identified ID number that allows for longitudinal tracking across the whole period. We also have information on the largest source of employment for every individual 16 years old or older who is employed by an organization in the Region during the period, her gender and age range. Longitudinal traceability of employees is done with another set of de-identified unique IDs, so we can compute flows of employees moving between organizations across time.

We study organizations of 1000 or more employees. Due to the highly skewed nature of the size distribution—a statistical feature of the Stockholm Region as well as other economies, see e.g. (Axtell 2001)—there are relatively few such organizations in the dataset, an average of 84 over time. However, they concentrate approximately 37% of the employed population in the Region (an average of 314,000 employees). Understanding how their social space looks like gives insights into a dominating fraction of the whole economy. On the other hand, our methodological strategy of homogeneity analysis and employee flows works best if the number of observations is relatively small, due to the fact that visualization of homogeneity and flow community patterns becomes very complex—if not impossible—for a large number of points.²

¹ The name of the full database is LISA, for Longitudinal integration database for health insurance and labour market studies. More information on the source of the data can be found at the Statistics Sweden website: <http://www.scb.se/lisa-en/>.

² The lower limit at 1000 employees is motivated by the importance of large organizations in the size distribution and the possibility to make the number of organizations tractable by the homogeneity analysis method. However, the number in and of itself is arbitrary. We then performed two sensitivity analyses taking a size limit slightly less than 1000 employees, concretely at 950 employees, and a limit slightly above, at 1050 employees. The choice of variables and variable definitions were left constant. The results and conclusions reported in the study for the 1000 employee limit are robust with respect to the change in size limits. The only remarkable difference is that the sample with 950 or more employees, because it includes more organizations and employee flows, has more pairs of organizations separated by a large

A first step before performing the analysis is to define which variables to consider. We will keep the model as simple as possible and include only nominal variables. The Gifi framework and the implementation in the package *homals* allow for a number of extensions like using ranked variables (de Leeuw and Mair 2009). Table 1 shows a list of the variables used in our analyses.³ To begin with, most research on organizational dynamics and communities has focused on privately-owned organizations. Given that the Stockholm Region during the period of analysis has about half its population in the public sector, the interaction between public and private organizations is a factor that cannot be excluded. Consequently, a first variable to consider is the ownership sector, indicating if the organization is public or private. Additionally, the literature on organizational ecology has shown that different industries or organizational populations exhibit different birth, growth and death patterns (Hannan and Freeman 1989). So the sector ownership variable is complemented by a variable about the organization's main industrial activity, calculated as the activity with most employees. As a measure of activity concentration, we distinguish between organizations where the main activity occupies 75% or more of the employees, and otherwise.

Within large organizations, a further distinction can be made between the ones up to 10,000 employees and the larger ones. The reason for this distinction is that size effects have been found to be significant for many organizational demographic processes, from the disbanding of organizations to their survival in time (Carroll and Hannan 2000, Chap. 14). Thus, the size range constitutes another one of our categorical variables. The overall number of employees is a global measure of size affecting the whole organization. A local measure of size is given by the workplace density, which is the average number of employees per workplace. We code the variable into two categories: less or equal than 100 and larger than 100. Finally, it has been shown that the demography of organizations affects individual employee mobility, and that the effect is different depending on gender and age (Carroll and Hannan 2000, Sect. 19.4). In order to address these effects, we consider two demographic variables to at the workplace level: fraction of workplaces with at least 50% female employees (coded as 1 when at least 50% of the workplaces meet this criterion, and 0 otherwise), and fraction of workplaces with at least 50% employees

Footnote 2 continued

distance in homogeneity space. Consequently, there are more outliers in the distance distributions for the 950 employee limit, but the bulk of the distributions is qualitatively the same.

³ The use of variables in homogeneity analysis determines the coordinates in homogeneity space. In a sensitivity analysis, we checked how the results are affected when variables are excluded from the model. First, we test how the coordinates change when successively taking the variables in Table 1, in reverse order, i.e. from variable (7) to variable (1). The mean square deviation from the full model is then calculated for each successive stage. The results depend on the dimension we consider. In the first dimension there is little variation; one needs to take away five variables in order to have a relative mean square deviation larger than 10% of the original coordinate values. The second dimension of homogeneity space, however, is more sensitive, with relative mean square deviations oscillating from 23 to 110%. However, this situation takes place when we cumulated the exclusion of variables. If we instead take out one variable at the time and analyze the individual effect, the relative mean square deviations are under 1% of the full model for most variables as well as both homogeneity dimensions. The two variables that have a largest effect are the fraction of female- and young-dominated workplaces.

Table 1 Variables and its respective category values for homogeneity analysis

Variable	Description	Category values
(1) <i>sec</i>	Ownership sector	public, private
(2) <i>sniGr</i>	Main industrial activity	Manufacturing and mining, Energy and waste, Commerce and communication, Financial and company services, Education and research, Human health and social work, Cultural services Public administration and others
(3) <i>sniGr75</i>	Fraction of employees in main activity	< 75%, ≥ 75%
(4) <i>sizeR</i>	Size range by number of employees	Large (≤ 10,000), very large (> 10,000)
(5) <i>empWP</i>	Average number of employees per workplace	≤ 100, > 100
(6) <i>femFrac</i>	Fraction of female-dominated workplaces	< 50%, ≥ 50%
(7) <i>youngFrac</i>	Fraction of young-dominated workplaces	< 50%, ≥ 50%

34 years old or younger⁴ (coded as 1 when at least 50% of the workplaces meet this criterion, and 0 otherwise).⁵

3 Methods

3.1 Homogeneity analysis

Homogeneity analysis is a descriptive dimensionality reduction method for categorical variables, meaning that it attempts to capture variation patterns in the data by representing the information in a structure that has considerably less dimensions than there are observations in the dataset. Homogeneity analysis belongs to the Gifi family of methods (Michailidis and de Leeuw 1998) that builds on and

⁴ The 34 year-limit corresponds to the classification in the Labor Force Surveys (*Arbetskraftsundersökningarna*, AKU) by Statistics Sweden. More information available at the website: <http://www.scb.se/am0401-en>.

⁵ We checked for the sensitivity of the model with respect to the choice of threshold values for some of the categorical variables. Changing the fraction of employees in main activity from 75 to 50% has very little effect: relative mean square deviations of less than 1% for both homogeneity dimensions. A change in the average number of employees per workplace from 100 to 50 impacts then first homogeneity dimension by 0.2% and the second by 4%. Finally, changing the thresholds in the fractions of female- and young-dominated workplaces from 50 to 75% results similarly small effects: 0.1 and 0.2% respectively for the first dimension, and 0.9 and 12% for the second dimension. Again, the second dimension of the variable for young-dominated workplaces is the most sensitive, but overall not as much as to alter the results significantly.

extends Multiple Component Analysis (MCA), the method pioneered in Bourdieu (1984).

Let us summarize how this method works. Traditional regression models describe—or explain as it is usually put—variation in the data by minimizing the residual of a given function with respect to the data, or equivalently by maximizing the likelihood of a given model to represent the data. Homogeneity analysis works by minimization of a loss function that depends on the categorical values taken by the observations. Without going into technical details, the goal here is to find values for both observations and category values—called object scores and category quantifications respectively—that minimize the loss function under the restriction posed by the input data. The resulting values are coordinates in an N -dimensional space where the dimensionality is usually 2 or 3. Following the logic outlined in the Introduction, this method is à la Bourdieu so to speak, because it does not use the categorical variables as coordinate axes, but rather lets the similarity in the data decide the placement of observations and category values.

An illustration of the output of a two-dimensional homogeneity analysis on our data for 1 year is sketched in Fig. 1, upper panel. Both organizations and the values of a given categorical variable (taken two values, blue and green in the example) are placed in the space. Each organization takes one and only one value per category, and thus a distance between organizations and category variable values can be determined. One of the most interesting features of this method is the possibility to interpret the estimations geometrically, and use the resulting space as a proxy for social space or at least one aspect of social space. In particular, the loss function has a nice geometrical interpretation as the sum of the square distances between all observations and all categories. So minimizing the loss function is a global optimization condition equivalent to finding the arrangement of points in the plane where the sum of these distances is minimal.

Additionally, there is the possibility to see how the space changes in time and therefore to characterize the dynamics or the organizational population from the point of view of the evolving social space. Here we should remind ourselves that our social space is structured à la Bourdieu, meaning that a change in social space follows from a change in the units (i.e. organizations) and unit categories that span it.

For the implementation of our homogeneity analyses we use the R package *homals* (de Leeuw and Mair 2009).

To conclude, let us mention that there are other methods closely related to homogeneity analysis. Multidimensional scaling (MDS) is a related technique that relies on writing down a dissimilarity function for the data. The loss function has another form where differences in dissimilarity matter, unlike the loss function in homogeneity analysis that looks at the norm of the differences. These results in MDS not having the same geometrical interpretation of space coordinates as homogeneity analysis, so our chosen method provides a more direct and useful interpretation. Michailidis and de Leeuw (1998) state that certain forms of cluster analysis in the sense that one tries to identify groups of observations based on similarity along categorical variable values, closely related to the idea in homogeneity analysis that similar observations should be placed in close coordinates.

Fig. 1 Homogeneity analysis output. Upper panel: organizations (circles in the figure) are placed on a 2D space. The more similar two organizations are along their categorical variable values, the closer to each other they feature in plot and the smaller the social distance. Category values also get coordinates in homogeneity space (squares in the figure). Organizations can take one and only one category value. In the illustration, there is one category with two category values, blue and green. Lower panel: network communities embedded in social space. This illustration expands upon the illustration in the upper panel. Here, we see additionally that organizations are linked by the employees changing jobs between them. Note how the network extends outside the space, since large organizations are linked among themselves as well as with smaller organizations. Groups of organizations that have considerably more movement edges within the group than between constitute a network community, also shown in the figure. (Color figure online)

3.2 Network community structure

One approach to characterizing the dynamics of organizational communities and their social space could be to look for connectivity structures in the links between organizations. In our case, the network is build out of employee movements

between organizations. Within the field of network science, the development of algorithms and procedures to detect and track down *network communities* constitutes an expanding field of research; see Fortunato (2010) for a relatively recent review and Surendran et al. (2016) in particular for a geometric approach. A network community—hereafter community—is a subset of organizations (nodes in the network) such that the links within nodes in the community are more prevalent than the links to nodes outside of it.

Our intention is not to engage in a comparison between the myriad of community detection algorithms out there, but rather to use network communities as a further dimension of social space structuring. Therefore, we focus on the widely used Louvain community detection algorithm (Blondel et al. 2008), implemented in `python`. It is a fast modularity-based algorithm that returns a so-called dendrogram, i.e. a hierarchical arrangement of communities that are successively aggregated to form higher-level communities. We can see an illustration of the interplay between the homogeneity space of organizations and network communities in Fig. 1, lower panel. Note that network communities are constituted both by large organizations—the ones in the homogeneity space being estimated—and other smaller organizations as well.

4 Results

4.1 Similarity dynamics in social space

4.1.1 Homogeneity analysis

We work with the two-dimensional space solution. Higher dimensions are possible, but the visual interpretation gets trickier in that case and often boils down to projecting the points into 2D planes anyways. The plot in Fig. 2 shows an example of how the output of homogeneity analysis for a particular year can be visually represented to conform a social space. The plot features the observation scores representing positions of organizations in social space. Positions corresponding to observations that have the same value of a given categorical variable (here public or private ownership sector) are shown in the same color. Note that a point may represent more than one organization, because two organizations that have the same combination of categorical variables will get the same scores. We furthermore observe that organizations belonging to the same sector tend to cluster in space. Recall that the minimization that results in the scores for each observation and category value is global (all data is used simultaneously) so the clustering we observe is not just a consequence of the sector variable, but of the overall variation in the data. A qualitatively analogous picture can be drawn for other years and categorical variables.

4.1.2 Category time evolution

The dynamics of these large organizations are very complex to describe. However, organizations can be grouped along various categories, and the time evolution of the

Fig. 2 Example of homogeneity analysis in a time point (year 1992). The axes correspond to the two dimensions specified by the dimensionality reduction in homogeneity analysis. Observation scores shape by the organization's ownership sector and color by main industrial activity. Quantifications for the corresponding category values also shown in the plot

whole category can give insights into the evolution of the organizations possessing a certain category value. It is here when having a method that provides a simultaneous geometrical interpretation for both organizations and category values comes is very useful. We first need to define a social distance. Two organizations i, j have their corresponding coordinates in homogeneity space, given by the pairs (x_i, y_i) and (x_j, y_j) respectively. Then the social distance is simply the Euclidean distance in two dimensions⁶:

$$d(i, j) = d(x_i, y_i, x_j, y_j) = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}. \quad (1)$$

The time evolution of category quantifications—sometimes called category centroids—for all dichotomous categorical variables is shown in Fig. 3, upper panel. The case of main industrial activity is omitted for clarity, as it would have

⁶ A note on null distances. As mentioned before, organizations having the same combination of variable values will end up in exactly the same homogeneity space coordinates. As a consequence, the social distance between them will be exactly equal to zero. The fraction of organizations occupying unique positions decreases over time, from 71% in 1990 to 27% in 2003. Among the combinations with the higher number of organizations per position are organizations in the public sector in health activities, with an average of 18.7 organizations per position. Thinking in terms of pairs of organizations, an average of 3.2% of all the pairs with no movement edges correspond to organizations in the same position, while the same is true for an average of 2.6% of the between-community edges and a 30.4% of the within-community edges. However, our results are not driven by the presence of these zero-distance pairs.

been necessary to show all possible combinations of pairs of industrial activities. We can see that the distances for most variables are relatively stable over the whole period, indicating that the corresponding variable values do not change on average their similarity. For instance, the distance between organizations having large concentration of employees in the same activity versus not having a large concentration (variable *sniGr75*) does not change much.

Exceptions are the size range variable *sizeR* and the variable on dominating fraction of young employees *youngFrac*. Large and very large organizations become overall more dissimilar as time elapses, with a stability period after the beginning of the 1990s and a reduction towards the beginning of the 2000s. The distance between organizations dominated by young people and others increases substantially at the beginning of the period and then decreases to remain stable. These drastic changes can be related to the Swedish crisis at the beginning of the 1990s, characterized by high unemployment and the progressive shrinking of large private organizations (Bergmark and Palme 2003).

The evolution of the distance between categories is informative, but equally interesting is to see how the distance changes look over time. The simplest change analysis is to look at the change in distance from 1 year to the next one, so-called *growth rate* (Stanley and Amaral 1996), defined for two objects i, j as the logarithmic ratio of their successive distance values:

$$r_d(i, j; t_0, t_1) = \log_{10} \left(\frac{d(i, j; t_1)}{d(i, j; t_0)} \right), \quad (2)$$

where t_0, t_1 stand in the usual convention for a reference year, say 1995, and the next year, in this case 1996. This rate is equal to zero when there has been no change in distance, positive when the categories have moved apart—i.e. become more dissimilar—and negative in the opposite case when they have come closer in social space. The results can be seen in Fig. 3, lower panel. We observe a clear divide between two variable groups: activity concentration and employee dominance (*sniGr75*, *femFrac*, *youngFrac*) tend to have positive growth, i.e. the distance between category values increases, in line with our previous observations. The other values have negative growth; organizations with different category values become more similar over time. In 1998 all variables reach a stall point with very little growth. Interestingly, this year corresponds roughly to the end of the recovery of the economy after the crisis at the beginning of the decade.

4.2 Communities in social space

Organizations do not exist in isolation; they are linked with one another in various ways, for example by the employee mobility between organizations. Here we consider the movement edges connecting the large organizations in the sample and the rest of the population in the Stockholm Region. Furthermore, we take into account the community structure of this network and use the Louvain algorithm to find and aggregate the communities up to the third level. It should be noted that for the purposes of detecting communities, the directionality of the movement edge is

Fig. 3 Time evolution of distance variables in homogeneity analysis. Upper panel: distance between category centroids of dichotomous variables. Lower panel: rate of change of distance between category centroids of dichotomous variables, according to Eq. (2). Of special interest is the year 1998 where the all rates of change are close to zero; this coincides roughly with the end of the recovery period from the crisis at the beginning of the 1990s. *Variables*: *sec*: ownership sector, *sniGr75*: fraction of employees in main activity, *sizeR*: size range (number of employees), *empWP*: average number of employees per workplace, *femFrac*: fraction of female-dominated workplaces, *youngFrac*: fraction of young-dominated workplaces

not considered, i.e. only the existence of a link between two organizations is relevant, not its direction. This allows us to see if the employee flow patterns within the same community differ from the patterns between communities.

4.2.1 Size and social distance: gravitational laws

In the Introduction, we took up an example illustrating how social interaction decays with geographical distance. This is a particular instance of a so-called *gravity law* from spatial statistics of population models (Erlander and Stewart 1990). Over the years, researchers have found other similar regularities in geographically bounded flows in social networks, see (Barthélemy 2011) for an overview. Some of these regularities take the functional form

$$F_{ij} = \frac{S_i S_j}{d(i, j)^\alpha}. \quad (3)$$

The case reported in (Zipf 1946) for movements of people between cities corresponds to setting the exponent $\alpha = 1$. Traffic flow in highways between cities and inter-city communication have been shown to decay with an exponent $\alpha \approx 2$ (Barthélemy 2011). Some of the functions are more complicated extensions of Eq. (3), but the one presented here already captures different kinds of phenomena.

But would the behavior for social distance be comparable to the geographical distance case? Our check for the gravity law is shown in Fig. 4, where we plot, for each pair of organizations over the whole period, the total flow of employees times the distance against the product of the respective organizational sizes (measured in number of employees).

Note that the flow is the total flow, namely the sum of employees going in both directions and not the net result. We also show reference slope lines corresponding to the cases $\alpha = 0.5, 1, 2$. We can observe that employee movement edges between organizations of 10,000 employees or less exhibit a random pattern with average slope close to $\alpha = 0$, regardless of the community structure. So, no apparent correlation is observed here. The case of within community large to very large edges is somehow similar, with a slightly positive slope. Unlike the between community case which features a slope close to $\alpha = 0.5$. If we replace the exponent in Eq. (3) we see that the flow decays slower with social distance than in the baseline case $\alpha = 1$. Thus, employees can reach our further in social space than they do in physical space. Edges between very large organizations have a certain slope, but the data points are too few to make any meaningful claim.⁷

4.2.2 Distribution of movement edges

As named in the Introduction above, homophily is one of the most robust empirical findings in social network analysis. On the other hand, it is assumed that social entities are connected to a higher degree to other entities in their vicinity in social space. So people should have on average higher contact frequency the lower the distance in social space. We can check if this statement holds for our population of large organizations in the Stockholm region. In Fig. 5 we show the time evolution of the distribution of distances computed according to Eq. (1), divided by whether we

⁷ Essentially the same pattern is observed if the data are broken down by year.

Fig. 4 Check for regularities of the ‘gravity law’ form according to Eq. (3). Given two organizations i, j separated a social distance d_{ij} and exchanging a total labor flow F_{ij} , the product of the flow and the distance is plotted against the product of the respective sizes S_i, S_j . Data points broken down by edges between/within communities, and between organizations of different size ranges (large $\leq 10,000$ employees, very large $> 10,000$ employees). OLS fits to the data shown in solid lines. Reference slope dashed lines corresponding to exponent values $\alpha = 1/2, 1, 2$. All years pooled together

observe a movement edge that year or not. We see that our hypothesis is in fact confirmed: the distribution for pairs of organizations with movement edges are consistently shifted towards lower distances in our space. The difference is more pronounced at the beginning of the period.

We performed a Chi square test for comparison of two density functions of the same number of bins according to (Press et al. 1992, Sect. 14.3) in order to see if the distributions are statistically different.⁸ Table 2 shows the result of the test. As we can see from the p-values, the distributions are statistically significant for all years. The p values proved to be quite insensitive to the number of bins.

⁸ Given a two histograms with 25 bins each, and with E_{1i}, E_{2i} the number of observations in the i th bin of the first and second histogram respectively, the Chi square statistic is constructed as $\chi^2 = \sum_i (\sqrt{E_2/E_1} E_{1i} - \sqrt{E_1/E_2} E_{2i})^2 / (E_{1i} + E_{2i})$, where $E_1 = \sum_i E_{1i}, E_2 = \sum_i E_{2i}$.

Fig. 5 Distribution of distance in homogeneity space, categorized by existence of a movement edge between the organizations, over time. Euclidean distance computed according to Eq. (1)

Table 2 Chi square test for distribution comparison, corresponding to Figs. 5 and 6

Year	Figure 5		Figure 6	
	χ^2	p value	χ^2	p value
1991	180.179	0.000	388.446	0.000
1992	229.210	0.000	328.234	0.000
1993	248.270	0.000	256.917	0.000
1994	187.105	0.000	309.726	0.000
1995	222.987	0.000	375.618	0.000
1996	244.503	0.000	292.458	0.000
1997	271.335	0.000	446.754	0.000
1998	274.630	0.000	327.959	0.000
1999	242.022	0.000	456.776	0.000
2000	261.282	0.000	627.683	0.000
2001	191.156	0.000	492.254	0.000
2002	289.252	0.000	430.086	0.000
2003	308.765	0.000	508.830	0.000

4.2.3 Community stability and social distance growth

The case of edges within versus between communities is shown similarly in Fig. 6. A Chi square contrast was performed in this case as well. The results, also shown in Table 2, indicate again that the distributions are statistically different; organizations belonging to the same communities are on average closer in homogeneity space. This reinforces the previous observation: organizations closer in social space are both more likely to be linked by employee movements, but also more likely to

Fig. 6 Distribution of movement edges within and between communities over time

belong to the same network community. One result is related to the other but not intuitively deductible from it.

The distribution of edges over time provides an overall sense of how social distance behaves between and within network communities. A related question, presented in the Introduction, concerns the stability of organizational communities as a whole. A first approximation to that is to look at how the social distance between pairs of organizations changes from 1 year to the next. Similarly to what we did for groups of organizations, we can use the growth formula in Eq. (2) and study the probability density functions. The analyses are shown in Fig. 7. The probability density functions are separated by type of edge: no edge, employee movement edge between communities and edge within communities. We can see that the patterns for pairs of organizations with no movement edge and for edges between communities are essentially equal. The resulting distribution resembles a so called *double exponential or Laplace* distribution, a robust recurring feature in growth processes of many kinds; see Mondani et al. (2014) for a review of applications and models. A characteristic of this statistical pattern is that change in distances is most of the time very small, but occasionally some abrupt change occurs, represented by the heavy tails in the distribution. A case of random distance fluctuation would lead to a Gaussian distribution, and that is roughly the pattern we observe for within-community social distance growth.

The distributions are furthermore broken down by time ordering of the edge, meaning simply that the data points are grouped according to the existence of an edge the year before and the current year. For example, $edge(t_0, t_1) = No \rightarrow Yes$ translates to all pairs of organizations during all years where an edge did not exist 1 year but existed the year after that. The motivation behind this was mentioned in Sect. 1: to try and check whether the similarity dynamics is the result of homophily or social influence, at least in a 1-year time window. Since there are essentially no differences between the distributions, we conclude that in the short term there is no way to tell if similar organizations become linked, or if linked organizations become

Fig. 7 Conditional probability density function of distance growth rates between pairs of organizations, according to Eq. (2). Distributions computed by type of organizational pair: no movement edge, edge between network communities and edge within communities. All years pooled together

more similar, or both. See “[Appendix](#)” for an additional check related to initial distance.

5 Discussion

5.1 Conclusions

In this study, we used Swedish longitudinal register data and a combination of homogeneity analysis and network community detection to study the effect that similarity in organizational properties has on the likelihood of interaction between organizations. For that, we characterize the dynamic social space of large organizations in the Stockholm Region for the period 1990–2003. Our conceptualization of social space is relational, drawing inspiration from the work of Bourdieu (1984), as opposed to a space of absolute coordinates à la Blau.

We performed a dynamical homogeneity analysis in two-dimensional space. Formulating the problem within this framework allows for a geometrical interpretation of similarity, thus promoting the problem to a geometrical problem. We analyzed similarity dynamics of groups of organizations by studying the time evolution of the distance between category values and found that most categorical variables are quite stable over time, with the exception of organizational size range

and fraction of workplaces dominated by young people. We further associated these variations with the Swedish crisis of the 1990s and its recovery phase.

Distance in social space was shown to affect the likelihood of a movement link between two organizations: linked organizations have a lower average distance than organizations not sharing a movement edge. This is in line with the usual theoretical assumption about proximity in social space favoring the emergence of social relationships.

We then studied the application of social distance-based analyses to organizational communities defined by inter-organizational employee flows. Our results show that the dynamics of employee movements within organizations belonging to the same community are in general qualitatively different from the dynamics of flows between organizations in different communities. Pairs of organizations belonging to the same network community are closer in homogeneity space than pairs of organizations in different communities. An analysis inspired by spatial statistics in geographically bounded network flows showed that the flow of employees between organizations in different communities exhibits a so-called gravity law with an exponent that decays more slowly than observed geographical networks, meaning that employees reach out regions of social space further than of physical space.

The result that organizations reach out further in social space than in physical space could be due to different reasons. For instance, the cost of reaching out could be higher in geographical space, since it requires to physically transport goods and services. Organizational similarity lowers cost of information and other exchanges, and part of this is reflected in the geometry of homogeneity space. One could also imagine a thought experiment where the cost of reaching out in space is severely reduced. In the light of the slower decay of interaction in social space compared to physical space, a drastic cost reduction could have a larger impact in strengthening the interactions mediated through social space (like information exchange), thus weakening geographical boundaries.

Finally, we studied the relative change in social distance between pairs of organizations. The resulting growth distributions for non-linked organizations as well as for organizations belonging to different communities feature a heavy-tailed Laplace statistical pattern similar to the one found in numerous other growth processes in natural and man-made systems, while the within-community social distance growth is more random and Gaussian like. To the best of our knowledge, this is the first time this result is shown in this context.

Additionally, there are no signs that homophily as a mechanism for similarity dynamics is preferred to social influence, at least not the short term. These findings may have implications for the understanding of how an aggregation of random local fluctuations inside a unit like a community or a city, coupled with heavy-tailed fluctuations between units, can give rise to the statistical growth patterns.

5.2 Further research

Our exploration of the social space of organizations is far from exhaustive, and there are several avenues to pursue further research. Theoretically, one can imagine

further developing our homogeneity analysis into a fitness landscape model for organizational selection, something more in the spirit of McPherson and Ranger-Moore (1991) named in the Introduction. Regarding network communities, further analyses can be carried out, for instance, in order to focus on the evolution of the communities themselves and their change in time.

As mentioned on the Introduction, social space may very well consist on different aspects interplaying with each other. On that regard, we have explored how networks can be used to capture a relational aspect of dynamics in social space, in particular having to do with organizational communities. Another potentially interesting network aspect would be to explore how constrains in social space interact with social networks to facilitate the generation and accumulation of social capital, thus opening up the possibility to bridge the gap between the literature on social space models and the work on networks models of social capital (Burt 2000).

These developments can help identifying process regularities and lead to a better understanding of the mechanisms underlying the dynamics of interactions between organizations. As the analyses of change in social distance over time show, it might be possible to track macroeconomic phenomena by looking at the changes in social space, and maybe ultimately provide indications on e.g. how long a phase in the economic cycle will last. Along that line, to study the change and growth of distance in social space may offer a new tool to extract information about the whole system of organizations, and is applicable as well to any other context featuring a group of interacting social units.

On the methodological side, other representations of homogeneity analysis are relevant. As seen in Fig. 1, upper panel, the output with coordinates for observations and category variable values can be analyzed as a planar bipartite network, i.e., a network composed of two types of nodes—organizations and category values—where edges only link nodes of one type with nodes of the other type (Michailidis and de Leeuw 1998). This will give rise to a potentially very interesting combination of spatially embedded two-mode network dynamics. Furthermore, it is conceivable to postulate a social space that has a relative coordinate framework (à la Bourdieu) as we used here, embedded in a higher order absolute space (à la Blau). That would allow studying changes in the structure of the space as different from the changes in space geometry caused by changes in the social units populating it.

Finally, there are other related methods one may want to consider. Homogeneity analysis relies on diagonalization and eigenvectors to find the coordinates of observations and categories in the two-dimensional space. This is also the case of various network centrality measures based on eigenvector centrality (Bonacich 1972). One can therefore complement our network analyses with network measures. Lastly, there are other ideas leading to proxies for social space, and methods associated with them. For instance, within networks, there exists the *latent space* approach (Hoff et al. 2002). The latent space model, like homogeneity analysis, is à la Bourdieu and requires the dimensionality to be set from the start. In this case, the error in a regression model is given a geometrical interpretation.

Acknowledgements The author would like to thank Peter Hedström for assistance with the database, Mark S. Handcock for recommending the `homs` package, and Fredrik Liljeros for comments on earlier versions of this paper.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

Appendix: Scaled distance change distribution

In our previous analyses, we have shown that the conditional probability density function of distances between organizations not linked as well as linked but in different network communities exhibit heavy tails alike many known growth processes in natural and man-made settings including organizations (cf. Fig. 7). Within the literature on growth processes in general and on organizational growth in particular, it is customary to do an additional check by breaking the distribution down by reference distance (Amaral et al. 1997). The rationale behind this is to see whether the observed heavy-tailed distribution is merely a superposition of other types of distributions (say a Gaussian distribution) or if it on the contrary holds for all distance ranges D . To make the different ranges comparable we scale the values with

Fig. 8 Probability density function of scaled distance growth rate, according to Eq. (4). Distributions by initial distance bin

$$r_d(t_0, t_1)_{\text{scal}} = \frac{\sqrt{2}(r_d - \bar{r}_d(D))}{\sigma_{r_d}(D)}, \quad (4)$$

where $\bar{r}_d(D)$ is the mean distance growth rate for all pairs of organizations having a reference distance in the range D , and $\sigma_{r_d}(D)$ is their corresponding standard deviation. This procedure is similar to the standardization of a normal distribution. The resulting distributions are shown in Fig. 8. We see that the statistical pattern observed in the original analysis holds partially, mostly in the case of middle distances (0.01–0.1), green in the plot. This is particularly so for the distributions with no edge and between communities. The distributions within communities are qualitatively different from the rest, with heavy tails only for the decreasing side. Furthermore, the very small distances do not collapse on top of the other curves for the cases with edges, and instead looks like a highly-peaked distribution, a possible indication that the growth dynamics of very similar organizations are somehow different from the rest.

References

- Aldrich HE, Ruef M (2006) Organizations evolving, 2nd edn. Sage, Los Angeles
- Amaral LAN, Buldyrev SV, Havlin S, Maass P, Salinger MA et al (1997) Scaling behaviour in economics: the problem of quantifying company growth. *Physica A* 244(1–4):1–24
- Astley WG (1985) The two ecologies: population and community perspectives on organizational evolution. *Admin Sci Quart* 30(2):224–241
- Axtell RL (2001) Zipf distribution of U.S. firm sizes. *Science* 293(5536):1818–1820
- Barthélemy M (2011) Spatial networks. *Phys Rep* 499(1–3):1–101
- Bergmark K, Palme J (2003) Welfare and the unemployment crisis: Sweden in the 1990s. *Int J Soc Welf* 12:108–122
- Blau PM (1977) A macrosociological theory of social structure. *Am J Sociol* 83(1):26–54
- Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. *J Stat Mech* 2008(10):P10008
- Bonacich P (1972) Factoring and weighting approaches to status scores and clique identification. *J Math Sociol* 2(1):113–120
- Bourdieu P (1984) *Distinction: a social critique of the judgement of taste*. Harvard University Press, Cambridge
- Bourdieu P (1985) The social space and the genesis of groups. *Theor Soc* 14(6):723–744
- Bourdieu P (1987) What makes a social class? on the theoretical and practical existence of groups. *Berkeley J Sociol* 32:1–17
- Burt RS (2000) The network structure of social capital. *Res Organ Behav* 22:345–423
- Butts CT, Acton RM, Hipp JR, Nagle NN (2012) Geographical variability and network structure. *Soc Netw* 34(1):82–100
- Carroll GR, Hannan MT (2000) *The demography of corporations and industries*. Princeton University Press, New York
- Erlander S, Stewart NF (1990) *The gravity model in transportation analysis: theory and extensions*. Topics in Transportation, VSP, Utrecht
- Fortunato S (2010) Community detection in graphs. *Phys Rep* 486(3):75–174
- Hannan MT, Freeman JH (1977) The population ecology of organizations. *Am J Sociol* 82(5):929–964
- Hannan MT, Freeman JH (1989) *Organizational ecology*. Harvard University Press, Cambridge
- Hoff PD, Raftery AE, Handcock MS (2002) Latent space approaches to social network analysis. *J Am Stat Assoc* 97(460):1090–1098

- de Leeuw J, Mair P (2009) Gifi methods for optimal scaling in R: the package homals. *J Stat Softw* 31(4):1–21
- Leuthold H, Hermann M, Fabrikant SI (2007) Making the political landscape visible: mapping and analyzing voting patterns in an ideological space. *Environ Plan B* 34:785–807
- McPherson JM (2004) A Blau space primer: prolegomenon to an ecology of affiliation. *Ind Corp Change* 13(1):263–280
- McPherson JM, Ranger-Moore JR (1991) Evolution on a dancing landscape: organizations and networks in dynamic Blau space. *Soc Forces* 70(1):19–42
- McPherson JM, Smith-Lovin L, Cook JM (2001) Birds of a feather: homophily in social networks. *Annu Rev Sociol* 27:415–444
- Michailidis G, de Leeuw J (1998) The Gifi system of descriptive multivariate analysis. *Stat Sci* 13(4):307–336
- Mondani H, Liljeros F, Holme P (2014) Fat-tailed fluctuations in the size of organizations: the role of social influence. *PLoS ONE* 9(7):e1005271–9
- Montgomery JD (1991) Social networks and labor-market outcomes: toward an economic analysis. *Am Econ Rev* 81(5):1408–1418
- Press WH, Flannery BP, Teukolsky SA, Vetterling WT (1992) Numerical recipes in C: the art of scientific computing, 2nd edn. Cambridge University Press, Cambridge
- Stanley MHR, Amaral LAN (1996) Scaling behaviour in the growth of companies. *Nature* 379(6568):804–806
- Stern C (1999) The evolution of social-movement organizations: Niche competition in social space. *Eur Sociol Rev* 15(1):91–105
- Surendran S, Chithraprasad D, Ramachandra Kaimal M (2016) A scalable geometric algorithm for community detection from social networks with incremental update. *Soc Netw Anal Min* 6(1):90
- Zipf GK (1946) The p1 p2/d hypothesis: on the intercity movement of persons. *Am Sociol Rev* 11(6):677–686

Hernan Mondani is a sociologist interested in modeling social organizing. In particular, he studies the statistical properties arising from organizational dynamics, and their microscopic origins from individual-level interactions. He draws heavily on tools from the frameworks of complex networks, social network analysis, complex dynamical systems, organizational ecology and statistical physics.