Thomas Fackler and *Nadzeya Laurentsyeva*

Gravity in Online Collaborations: Evidence from GitHub

Growing importance of immaterial goods and increasing digitization have enabled virtual production processes and have made virtual teamwork possible. Modern digital technologies not only reduce communication costs, but also help create powerful environments with apparently no physical barriers for close collaboration or for exchange of knowledge and ideas. Does it mean that traditional obstacles, such as bilateral distance, country borders, language barriers or cultural differences do not matter in the virtual production processes? We try to answer this question by estimating the gravity model for collaborations on GitHub—the world's largest online platform for software development.

The gravity models are well established in the Economic literature and help to identify the determinants of bilateral trade in goods and services or of migration flows between geographical units. By applying the gravity model to an online setting, we can identify the determinants of virtual collaborations and compare them with those established in trade or migration literature.

Cross-city and cross-country code contributions are not only related to trade, but also to the literature on knowledge flows and knowledge production. Knowledge has been shown to be more localized than what would be expected from agglomeration effects alone (Jaffe et al. 1993). Furthermore, knowledge spillovers to other countries has been shown to take time (Hu and Jaffe 2003; Jaffe and Trajtenberg 1999), and the effect of international localization has turned out to be more robust than within-country localization (Thompson and Fox-Kean 2005).

Our results show that there is gravity in online collaborations on GitHub. The estimations suggest that it is weaker than in trade, but statistically significant, despite the fact that both the production process and the output of programmers are immaterial. The effect of distance between locations is non-linear, i.e., an additional kilometer decreases collaboration more when distance is low than when owner and committer are already far apart. This is in line with the idea that offline work and personal contact are still important and different modes of transport are used, such that moving from what may be a commuting distance to one that is usually traveled by plane changes the cost of an additional kilometer.

In addition, when distance is controlled for, traditional determinants of international trade such as language barriers and country borders matter for international code contributions, although here too the magnitudes of the effects are smaller than for trade.

CONTEXT AND DATA

GitHub is a platform for software development that was launched in 2007 and hosts a collaborative version control system. Projects can be started by individual users and companies. The repositories cover a wide variety of (mostly) software projects, some of which are aimed at other developers and some at a wider audience. GitHub allows users to have private and public repositories for the project's code. Our data contains only the latter. These public repositories are usually licensed under common open-source licenses such as the GNU General Public License.

To contribute to projects or create new ones, users have to set up an account and can provide their real name, location (usually city) and additional biographical information. Each project has only one owner. The owner may invite other users to contribute and become project members. Users can also initiate and contribute to a project before being invited (McDonald and Goggins 2013). Users who are not project members cannot only report issues but also suggest modifications to the code, which the project members can review and accept into the project.

In public projects, all of these activities can be observed by everyone. This makes collaborative software development a unique setting that gives researchers a detailed and, in terms of code, comprehensive view of worker interaction. Users' profile pages on GitHub show their contributions to different projects, while project pages reveal which users have contributed. Thanks to the version control system, the development history of a project is recorded down to the addition of each line of code. In addition to tools

Thomas Fackler

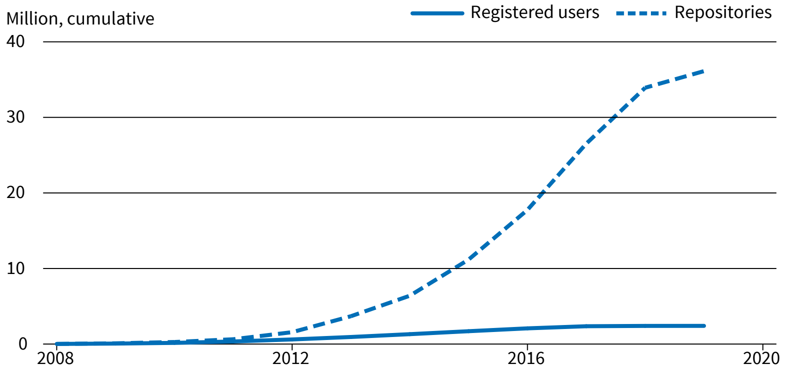
is an Economist at the ifo Center for Industrial Organization and New Technologies and a Postdoctoral Researcher at the Chair of Organizational Economics at the LMU Munich.

Nadzeya Laurentsyeva

is a Postdoctoral Researcher at the Chair of Organizational Economics at the LMU Munich and an Associate Research Fellow at the Economic Policy Unit of CEPS, Brussels.

Figure 1

Cumulative Number of Registered Users and Repositories on GitHub



Note: Only users with reported location; only repositories owned by users with reported location.
Source: Data from GitHub Torrent project (snapshot from June 2019).
© ifo Institute

for software development, GitHub also shares some features of social networks, giving users the ability to get updates about each other's activities, as well as watch projects and give "stars" to the ones they like. Motivations of open source contributors have been the subject of economic research and include paid work at software companies, career concerns (showcasing skills), as well as writing software for one's own needs or to help others (Belenzon and Schankerman 2008; Hergueux and Jacquemet 2015; Lerner and Tirole 2001 and 2005).

For this study, we mainly look at "push events," i.e., submissions of commits to a repository, and here in particular, the ones involving project owners and committers from different countries.

We use a snapshot from GitHub Torrents (Gousios 2013) and the GitHub Archive Dataset, as well as a Gravity dataset from CEPII. Both Torrents and Archive datasets provide a mirror of the GitHub public event stream from 2012 onward. Both are publicly available in the Google Cloud Platform. We use the two datasets in a complementary way. We take the event stream data from GitHub Archive as it is updated in real time and allows us to incorporate the most up-to-date activity data. We then merge the events with data on users (in particular, their reported geographic locations), which is available in the Torrents dataset. We use the latest available snapshot of GitHub Torrents from June 2019. Thus, our event data spans from 2012 to July 2020, conditional on the involved users (project owners and project committers) being registered on GitHub as of June 2019.

Our final dataset has several features. First, it contains the available information from *public* repositories only, as we cannot observe the activity of private projects stored on GitHub. Second, given our research question, we have to limit the data to events where we can identify the location of project owners and project committers. As Figure 1 shows, that leaves us with about 2.4 million registered users and about 36 million repositories.[1] Third, to focus on the collaborative work, we keep only those events on GitHub where a project committer is different from the project owner.

GEOGRAPHY OF THE ACTIVITY AND COLLABORATIONS ON GITHUB: DESCRIPTIVE DATA

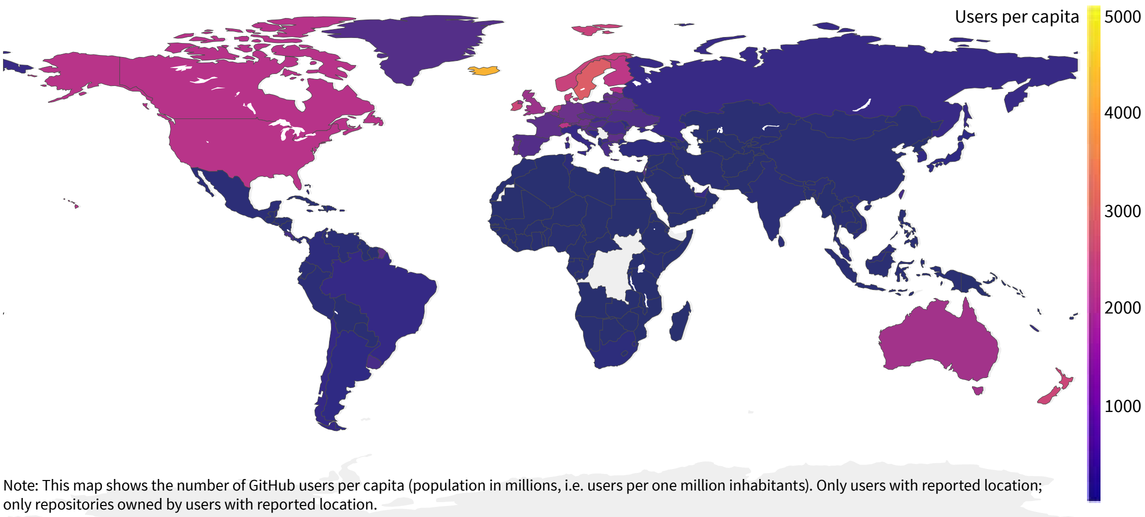
Since its start in 2007, GitHub has become popular with users around the world. Figure 2 shows the number of GitHub users in our data relative to a country's population (in millions). Overall, more advanced countries have a higher share of registered users. It should be noted that even though per-capita activity is highest in North America, Europe and Oceania, populous countries such as India and China have sizable user bases on GitHub as well.

The scatter plot in Figure 3 shows that the share of GitHub users per capita is highly correlated with

[1] In total, as of June 2019 there were 32 million registered users on GitHub and 125 million repositories.

Figure 2

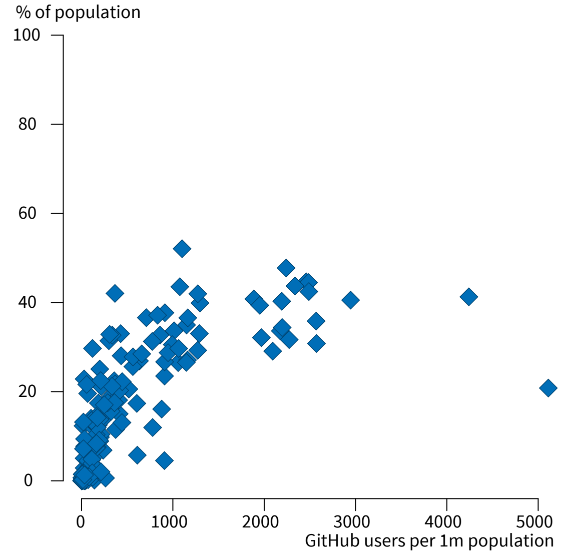
Number of GitHub Users per Capita



Note: This map shows the number of GitHub users per capita (population in millions, i.e. users per one million inhabitants). Only users with reported location; only repositories owned by users with reported location.
Source: Data from GitHub Torrents project (snapshot from June 2019).
© ifo Institute

Figure 3

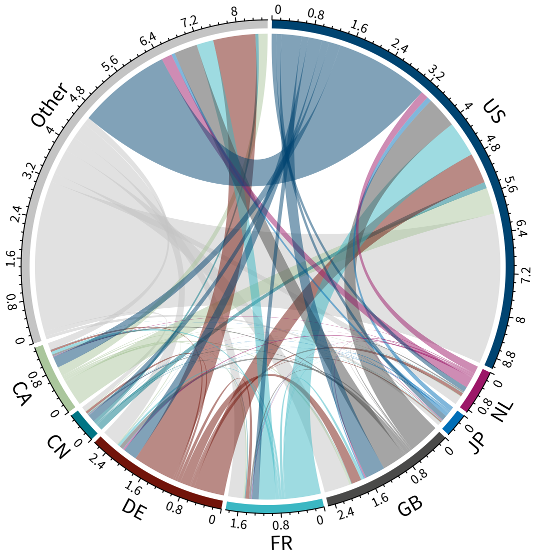
Broadband Subscriptions as a Share of Population and Number of GitHub Users per One Million Population



Note: This scatter plot shows the number of GitHub users per one million inhabitants of a country on the horizontal axis and the number of broadband subscriptions relative to a country's population on the vertical axis.
Source: Data from GitHub Torrents project (snapshot from June 2019); World Bank World Development Indicators 2018. © ifo Institute

Figure 4

Bilateral Flows on GitHub



Note: Number of contributions in millions. Only users with reported location; only repositories owned by users with reported location.
Source: Data from GitHub Torrent project (snapshot from June 2019). © ifo Institute

the number of broadband subscriptions per capita. Even though a slow Internet connection is technically sufficient for working on GitHub, broadband certainly helps, especially when other tools, such as video conferencing, are used for coordination purposes. Of course, a country's level of technological development correlates with both the share of programmers and the share of Internet users in a country. Our data shows that there is also a positive, although weaker, correlation between the share of information and communications technology in a country's exports and the number of users per capita.

Figure 4 depicts the flows of contributions between the eight most active countries on GitHub in our data in terms of international contributions (US, Great Britain, Germany, France, Canada, the Netherlands, China, and Japan), as well as to and from the set of all other countries. Within-country contributions are excluded. The circle shows how the international contributions between the illustrated countries are divided by committers' countries. The flows go toward the project owner's country.

The largest flow between two countries is from committers in France to projects whose owners are in the US (about 700,000 cross-border events), closely followed by commits from Great Britain (600,000), Germany (600,000) and Canada (500,000) to the US. The next-largest flow between countries is from the US (committers) to Great Britain (owner location), which is about two-thirds the size of the reverse flow. Among the countries shown, the top three countries by inflow of contributions to projects owned in the country are the US, Great Britain and Germany. The top three by "outflows" are the same countries, but Germany is in second place and Great Britain is third.

If the total of "outflows" (contributions to foreign projects) is divided by the total of "inflows" (contributions to local projects by foreigners), Germany has the highest ratio (about 2.4) and the US the lowest (0.6). This is interesting in view of the discussion about Germany's scarcity of technology start-ups relative to the US, despite the availability of local engineering talent. It is also in line with the political debate about Germany's export strength and American concerns about the trade balance, even though we are analyzing numbers that do not enter trade statistics. Japan and China, however, are the other two among the shown countries with a ratio smaller than one, despite their export strength.

ESTIMATION OF THE GRAVITY EQUATION

The gravity equation models bilateral interactions between geographic units where economic size and distance effects enter multiplicatively. In particular, the scope of interactions is positively related to partner size, which could be measured by GDP, income or population, and negatively related to bilateral distance. Such models have been used as a workhorse for understanding the determinants of bilateral trade flows for over 50 years, since being first introduced by Tinbergen (1962) − see Head and Mayer (2014) for a recent survey. They have also been widely applied to study the determinants of migration flows, see Beine et al. (2016) and Ramos (2017) for reviews of modelling approaches, and Mayda (2010) and Migali et al. (2018) for applications to international migration.

To estimate the gravity equation for collaborations on GitHub, we aggregate the data at a city-pair and year level. We further restrict our dataset to about 500 of the most active cities on GitHub (as proxied by the number of registered users as of June 2019).[2] These cities together account for over 70% of all commits by users with reported locations. We construct a strongly balanced panel dataset by forming all possible city pairs from our sample for a period between 2012 and 2020, which results in about 2.3 million observations.

We estimate several variations of the gravity model. Our baseline specification is the following:

$$cijt = \beta_0 + \beta_1 dij + \beta_2 X + \tau t + Eijt$$

c*ijt* is the number of collaborations between a city pair *ij* in a year *t*; we measure it by the number of contributions (commits to a project) done by users from a city *i* and submitted to a project owned by users from a city *j*. In our setting, direction matters: collaborations between city pairs *ij* and *ji* are treated as two observations. To make an analogy in terms of the trade and migration literature, we think of a city of committers as an origin (e.g., origin of service providers—exporters) and a city of the project owner as a destination (e.g., destination of services—importers). *dij* is geographic distance between two cities. We calculate it as the shortest path (in km) between cities, using their coordinates. *X* includes a vector of controls. We control for the number of users in origin and destination cities registered on GitHub as of a given year. In addition, we add a dummy for foreign country and a dummy for common language (for cross-border collaborations). Conditional on distance, these dummies capture the effects of state borders and language barriers. All the specifications include year fixed effects, and standard errors are clustered at a country-pair level to allow for correlations in residuals.

Table 1

Gravity Model for Collaborations on GitHub

Variables	(1) Contributions	(2) Contributions	(3) Contributions	(4) Contributions intensive
Distance	– 0.180*** (0.023)	– 0.167*** (0.037)		
1–50 km			– 3.038*** (0.443)	– 1.489*** (0.375)
50–100 km			– 4.372*** (0.121)	– 2.691*** (0.108)
100–300 km			– 4.931*** (0.122)	– 3.080*** (0.085)
300–700 km			– 5.072*** (0.123)	– 3.217*** (0.128)
>700 km			– 5.172*** (0.119)	– 3.342*** (0.094)
Users, destination	0.111*** (0.026)	0.111*** (0.026)	0.106*** (0.025)	0.305*** (0.039)
Users, origin	0.097*** (0.025)	0.097*** (0.026)	0.092*** (0.024)	0.204*** (0.041)
Foreign country		– 0.097 (0.117)	– 0.221*** (0.020)	– 0.427*** (0.040)
Common language		0.046** (0.018)	0.021** (0.010)	0.049 (0.082)
Observations	2,331,693	2,313,405	2,313,405	94,619
R-squared	0.132	0.135	0.253	0.264
Clusters	5184	5041	5041	2170

Note: The dependent variable is the number of contributions (natural logarithm + 1) between a given city pair. Column 4 presents results conditional on non-zero contributions in a city pair. In Columns 1–2: distance represents the length in km (natural logarithm + 1) of the shortest path between two cities. In Columns 3–4: we use dummies corresponding to different distance bins, where distance = 0 (same city) is the reference category. Economic size is proxied by the number of registered users in an "origin" city (city of a committer) and a "destination" city (city of a repository's owner). All specifications in-clude year fixed effects. Standard errors are clustered at a country-pair level.

Source: Own calculations.

In our baseline estimations, we take natural logarithms of our dependent and non-categorical independent variables. Therefore, we can interpret the coefficients of interest as elasticity. However, given that we have count data and many zero observations, for robustness, we estimate the regressions using zero-inflated Poisson method.

RESULTS

Table 1 presents our main results.[3] Columns (1) and (2) use a continuous measure of distance as the explanatory variable. The effect of geographic distance on online collaborations is negative and statistically significant with an estimated elasticity of 0.17–0.18. The magnitude of the effect is smaller compared to those established for trade (around 0.85 – 1) and slightly smaller compared to those found in the international migration literature (around 0.25). Yet, the effect is still sound, meaning that geographic distance matters even in virtual environments. Column (3) uses distance bins instead of a continuous distance measures to capture non-linear distance effects. The reference category corresponds to collaborations within the same city. The results highlight non-linearity in the distance effect and suggest that interactions on GitHub are substantially more likely to happen within the same city, i.e., between people who know each other personally and/or can collaborate in an offline setting. Beyond the distance of 100 km (roughly commuting distance), the effect stays at about the same level. Columns (2–3) also control for state borders and language. As in the trade and migration literature, conditional on distance, the state borders reduce virtual collaborations, while a common language slightly mitigates this negative effect. Column (4) focuses on the intensive margin and shows that geographic distance as well as state borders also matter for the intensity of collaborations.

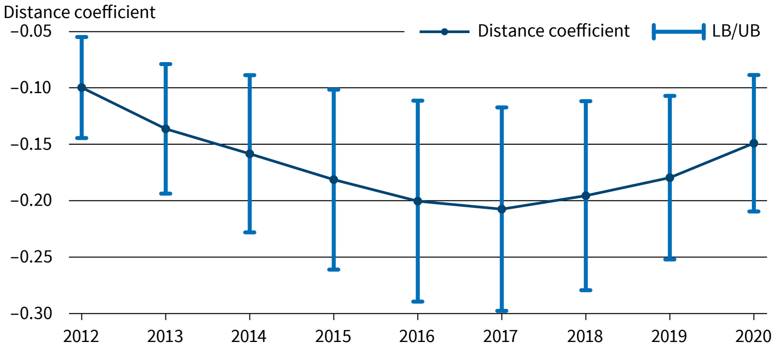
[2] We set a cutoff of at least 450 registered users per city as of June 2019, resulting in 511 cities.

[3] All our results are qualitatively robust to including city fixed effects and to an alternative estimation method with zero-inflated Poisson.

Figure 5 investigates whether the effect of distance on GitHub collaborations changed between 2012 (the launch of the platform) and 2020. Despite increases in Internet speed and online collaboration tools, our data suggests that the role of distance on GitHub did not substantially decrease over the last several years (if anything, it increased slightly between 2013 and 2017). The figure, however, is hard to interpret as the platform grew rapidly between 2013 and 2017. One possible explanation is that new users are more likely to start their collaborations locally or that recent user growth comes from professional users, who might be more likely to be co-located in offices than unpaid volunteers in open-source projects.

Figure 5

Distance Elasticity of Collaborations on GitHub in 2012-2020



Note: Distance elasticity is calculated by running separate regressions (same as the specification (2) in Table 1) for each year.

Source: Authors' calculations.

© ifo Institute

CONCLUSION

To summarize, results in Table 1 and Figure 5 highlight that standard barriers found to affect trade and migration flows also matter in a virtual environment. This is particularly interesting given that (monetary) search costs for a relevant project, technology or a potential partner on GitHub are zero. There are neither the usual "trade" costs, such as tariffs or quotas, nor any travel costs. Moreover, in a transparent setting such as GitHub, the information about the quality of a potential project or a contributor is easy to observe for all the actors. This finding is consistent with Singh and Marx (2013), who show that advances in communication technologies and lower costs of traveling hardly reduce the localization of knowledge over time.

There could be several explanations behind the effect of distance and country borders on GitHub. First, it could be driven by the motivation of programmers working on GitHub. If a programmer's main motivation to contribute to a certain project is career driven and if they consider mainly geographically close labor markets, they might focus their activity on local projects. Second, it is likely that personal contact and offline communication among co-workers matter even for online production processes. While GitHub offers infrastructure for virtual collaboration, certain problems (especially those related to the strategic development of a project) require personal interaction. Third, while software products and programming languages are relatively standard, substantial geographic differences in the contents, available technologies, and approaches to work are likely to exist, making projects from different cities and countries non-compatible. From a non-technical perspective, cultural differences could also play a role. For instance, Lyons (2017) uses data from an online contract labor market and shows that team organization improves outcomes when workers are from the same country. She argues that the effect is driven by easier communication among team members. Laurentsyeva (2019) uses GitHub data and provides evidence that political conflicts (which are completely exogenous to the functioning of GitHub) increase ingroup-outgroup biases among programmers from the affected countries and decrease cross-border collaboration.

Disentangling the exact reasons behind gravity in online collaborations using micro-level data from GitHub is a promising avenue for further research.

REFERENCES

Beine, M., S. Bertoli and J. F. H. Moraga (2016), "A Practitioners' Guide to Gravity Models of International Migration", *The World Economy* 39, 496–512.

Belenzon, S. and M. Schankerman (2008), "Motivation and Sorting in Open Source Software Innovation", *CEPR Discussion Papers* 7012.

Gousios, G. (2013), "The GHTorrent Dataset and Tool Suite", in *Proceedings of the 10th Working Conference on Mining Software Repositories,* IEEE Press, San Francisco, 233–236.

Head, K. and T. Mayer (2014), "Gravity Equations: Workhorse, Toolkit, and Cookbook", in G. Gopinath, E. Helpman and K. Rogoff (eds.), *Handbook of International Economics*, vol. 4, Elsevier, Amsterdam, 131–195.

Hergueux, J. and N. Jacquemet (2015), "Social Preferences in the Online Laboratory: A Randomized Experiment", *Experimental Economics* 18, 251–283.

Hu, A. G. Z. and A. B. Jaffe (2003), "Patent Citations and International Knowledge Flow: The Cases of Korea and Taiwan", *International Journal of Industrial Organization* 21, 849–880.

Jaffe, A. B. and M. Trajtenberg (1999), "International Knowledge Flows: Evidence from Patent Citations", *Economics of Innovation and New Technology* 8, 105–136.

Jaffe, A. B., M. Trajtenberg and R. Henderson (1993), "Geographic Localization of Knowledge Spillovers as Evidenced by Patent Citations", *The Quarterly Journal of Economics* 108, 577–598.

Laurentsyeva, N. (2019), "From Friends to Foes: National Identity and Collaboration in Diverse Teams", *CRC TRR 190 Discussion Paper* 226.

Lerner, J. and J. Tirole (2001), "The Open Source Movement: Key Research Questions", *European Economic Review* 45, 819–826.

Lerner, J. and J. Tirole (2005), "The Economics of Technology Sharing: Open Source and Beyond", *Journal of Economic Perspectives* 19, 99–120.

Lyons, E. (2017), "Team Production in International Labor Markets: Experimental Evidence from the Field", *American Economic Journal: Applied Economics* 9, 70–104.

Mayda, A. M. (2010), "International Migration: A Panel Data Analysis of the Determinants of Bilateral Flows", *Journal of Population Economics* 23, 1249–1274.

McDonald, N. and S. Goggins (2013), "Performance and Participation in Open Source Software on GitHub", in *CHI '13 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '13, Paris, 139–144.

Migali, S. et al. (2018), *International Migration Drivers,* Publications Office of the European Union, Luxembourg, https://publications.jrc.ec.europa.eu/repository/bitstream/JRC112622/imd_report_final_online.pdf.

Ramos, R. (2017), “Modelling Migration”, in L. Matyas (ed.), *The Econometrics of Multi-Dimensional Panels*, Springer, Berlin, 377–395.

Singh, J. and M. Marx (2013), “Geographic Constraints on Knowledge Spillovers: Political Borders vs. Spatial Proximity”, *Management Science* 59, 2056-2078.

Thompson, P. and M. Fox-Kean (2005), “Patent Citations and the Geography of Knowledge Spillovers: A Reassessment”, *American Economic Review* 95, 450–460.

Tinbergen, J. (1962), *Shaping the World Economy: Suggestions for an International Economic Policy,* The Twentieth Century Fund, New York.