# Does Our World "weigh" Less Right Now? The Gravitational Pull in a Scientific Collaboration Network is Getting Weaker with Time

Guy Kelman*
Hebrew University of Jerusalem
superk@lbl.gov

Moshe Levy
Hebrew University of Jerusalem
mslm@huji.ac.il

Eran Manes*
Jerusalem Institute of Technology,
msemanes@gmail.com

## Abstract

*We study the geographical patterns of scientific collaboration from a large sample of research papers and letters written by two authors that appeared in the magazine Nature over two sub-periods, before and after the popularization of Internet use. We report three results: First, the distance distribution of co-authors is fat-tailed, in agreement with other studies that find a gravitational law in collaboration networks. Second, in the later period the distance distribution dominates the range of commute-distance and beyond (>50km), which renders the city the atomic unit for statistical testing. Last, strong geographical clustering remains a major generative factor in this network. Assuming the universality of this law, we estimate the gravitational constant from the pull between scientists in the network. We find that this constant has decreased two-fold over the last three decades while the other coefficients remain stable. This may indicate that the gravitational constant absorbs changes in the environment that render distances easier to cross, namely a "lighter world"*

## 1. Introduction

Collaboration networks have been the focus of many studies in recent years [1-7]. Without exception, these studies have found that geography plays an important role in science, where the collaboration strength follows a gravitational law. Gravitational laws apply also to other complex networks such as social networks [13], mail exchanges [14], mobile phone communication [15,16], and blogs [15]. For example: in a study of inter-city telecommunication [8] the strength of communication was found to be inversely related to the square of the distance between the cities and positively related to the product of their population masses.

The prominent elements that gradually transformed our communication patterns include the invention of Internet-based communication tools such as Email, Social Networks, the reduction in airfare, and the collapse of the iron curtain.

Indeed, the Internet era has brought dramatic changes into everyday life, breaking territorial boundaries and enabling fast, cheap and reliable communication from afar, while allowing large-scale data flow between distant individuals at a single mouse click. the globalization of markets and the blossoming of cybermediaries and e-commerce attest to this "global village" transformative impact of the Internet.

Also, the Internet has profoundly changed many attributes of the scientific production process. Background literature is much easier to trace and retrieve, journal turnaround is quicker, and papers can gain immediate worldwide reach by becoming available online. The main questions in this paper are therefore - Has the nowadays casual means of communication transformed the way which scientists form collaborations? How about at long geographical distances? Last, what can be said about the quantitative and qualitative nature of this change?

In this paper we capitalize on the fact that the emergence of the Internet - an important means of long-distance communication - serves as a before-after experiment to study the responsiveness of collaboration networks to changes in communication patterns. Our analysis allows us to investigate whether or not the scientific community was brought closer together over the two periods before and after the popularization of Internet use. The results of this estimation lend strong support to the existence of a *gravitational law* in the network, wherein the strength of a link is proportional to the

* Present addresses: GK - Lawrence Berkeley National Laboratory. EM - Ben-Gurion University of the Negev

URI: https://hdl.handle.net/10125/64085
978-0-9981331-3-3
(CC BY-NC-ND 4.0)

product of the masses of its connecting nodes and inversely proportional to the square of the geographical distance between them. The gravitational constant of the collaboration network appears two-fold greater in the "before" sub-period, meaning it was harder to bridge distances back then. This we may attribute to changes in underlying communication media that were apparently brought over by the Internet and other concurrent changes in communication costs, such as the drop in airfare.

As a testing ground, we chose data on full papers and letters which appeared in *Nature* and were co-authored by two collaborators. We examined two sub-periods: 1984-1994 and 1995-2015. These correspond to the pre-Internet and the post-Internet breakthrough, respectively. The years 1993-1995, when e-commerce giants Amazon and eBay opened for business and the NCSA released its first version of Mosaic, later to be repurposed as Microsoft Internet Explorer, lend themselves as natural candidates for the cutoff year. After which the use of the Internet became global. We also tried different cutoff years in the range 1992-1999 to find that they did not profoundly affect the results (see Methods). We focus on papers that were written by two authors, mainly in order to avoid excessive noise emanating from unknown contribution weights and distance measures connecting more than two authors. The dataset we use is made of 3,718 letters and full research articles written in the journal *Nature*, which identify a total of 7,256 unique author affiliation addresses. Our choice to rely on papers appearing in *Nature* is grounded in the following two considerations: First, *Nature* is a general interest journal, which deems unnecessary the need to account for heterogeneity in publication standards and culture between different academic disciplines. Second, considered to be amongst the most prestigious academic journals, *Nature* adheres to high and meticulous peer-review and publication standards. Moreover, a publication in this journal may well translate into immediate impact on academic promotion. As such, articles in *Nature* are often the result of prolonged endeavors and long-standing collaboration that require efficient communication. On balance, therefore, the focus on *Nature* helps us avoid excessive noise and is consistent with the cost-effectiveness principle.

The rest of the paper is organized as follows: We present the results of this analysis in the Results section. Next, we discuss the findings in the Discussion section, and lay out the conclusions. Last, we detail our workflow in the Methods section.

## 2. Results

Table 7 presents a summary of the network analysis performed on the full sample of 25k papers written by any number of coauthors, and the down-sample to two coauthors. It is notable that the 2-author dataset represent a network below the percolation threshold. This means that the giant component does not compose roughly >90% of the network, namely that it is not "giant." We focus on network analysis of the full dataset and later discuss geographical implications on the 2-author subset for the reasons that were outlined above. Over the two periods, the number of papers has risen by 25%, and respectively the number of distinct authors (nodes) grew. Every co-authorship is a fully-connected clique of authors, and due to the nature of this design the number of links grew squared. Assortativity decreased with time and so did the diameter and average path length. This means that the network is easier to cross. Now, importantly, the clustering coefficient became smaller in contrast with the geographical clustering which became more appreciable. However, we note that this is an artifact of the growing teams in the sense that we are likely to find more authors connecting with high between-ness among 400 coauthors than among two.

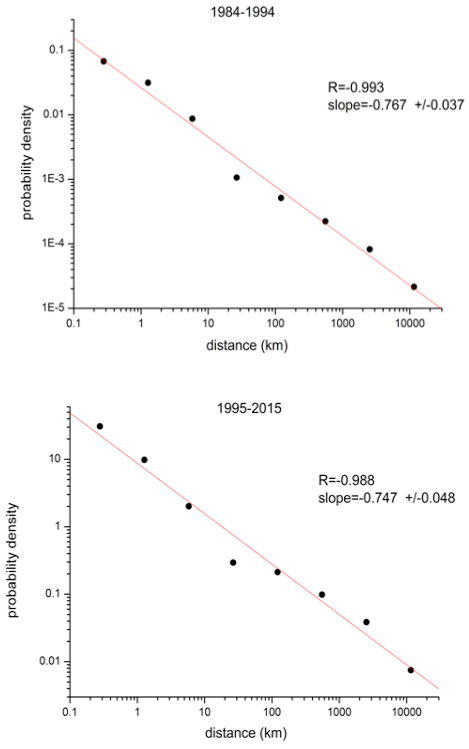


**Figure. 1**. Probability density functions for the two time frames on a double-logarithmic scale. The red line is a power-law reference fit.

Moving forward, Table 8 presents summary statistics on our downsampled papers and letters data. The number of paired collaborations is almost two-fold greater in the first period, with the overall number of papers being close to equal across the two periods of interest. This suggests that the number of papers with more than two co-authors is almost two-fold greater in the second period (in the full sample, the modal value of the number of authors is 2 in the first period and 3 in the second). The change in team size is also visible in the network analysis above. We speculate that this shift is motivated by the growing demand to facilitate teamwork and multi-campus projects that encourage global spread. Clearly there are competing explanations. One alternative is the increased competition in science and the resulting rise in tenure and promotion criteria, which forced academics into diversifying by collaboration. Another alternative is the "knowledge explosion" which has been a driving force for specialization, again fostering collaboration. The evolving trend in recent decades towards multiple authorship across many scientific disciplines is in fact well-documented in the literature. [9-12]

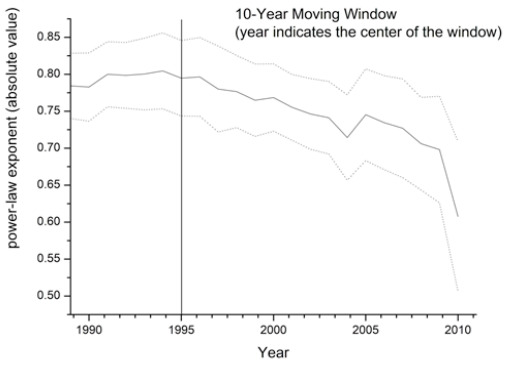


**Figure 2**. The slope of the distance-probability distribution (in absolute value) as a function of the year at the center of the 10-year window.

Throughout our analysis we focus on paired collaborations only. This choice is motivated primarily by the absence of a natural definition of distance between more than two collaborators. Further, it enables us to focus on simple, high impact projects rather than complex problems that require large consortia of resources. A similar approach was taken by Chandra et al. [3]

Figures 1 and 2 show the probability density of forming a paired collaboration with distance. The probability distributions exhibit similar decay across five orders of magnitude, with similar slopes.

While the slope is slightly smaller in the "after" period - the difference in slopes is not statistically significant.

To test the choice of the cutoff year for before vs. after we repeat the analysis in a 10-year moving window. Figure 2 displays the slope as a function of the year at the center of the 10-year window. It is evident that the slope *does* decrease systematically after 1995, indicating that the second period opened opportunity to form collaborations with less dependence on distance. Namely, that the reduction in communication costs driven mainly by the globalization of the Internet has been successful in spanning the geographical reach of scientific collaboration.

To gain further insight, Figure 3 plots the cumulative distance distributions of the two periods back to back. A first-order stochastic dominance of the first period is evident in the following sense: for any given distance $r_0$, the probability of co-authorship at distances $r \geq r_0$ is *at least as high* in the "after" period as it is in the "before" period. In other words, if $F(r), G(r)$ denote the "before" and "after" CDFs respectively, then $F(r) \geq G(r)$ for every $r \geq 0$.

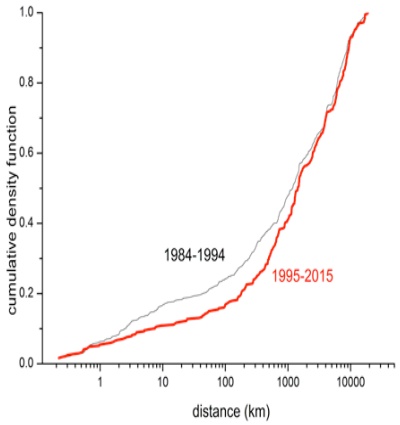


**Figure 3.** Cumulative distance distributions in the two time-frames

A more discerning interpretation of this figure is that the range 1-10km in the "before" period exhibits increased probability to find an occurrence of a pair of authors, whereas the range 10-50km does this for the "after" period. These distances are familiar as a range that spans the city's boundary and up to the commute distance, which haven't changed in the last 5 decades

[20]. For distances greater than 50km, there is greater chance to form pair collaborations in the "after" period.

Thus, we could sketch the following: For any given distance, there is a difference between the earlier and the later time frames. For the super-distant pairs we notice only a minute change in their probability to occur across the two periods. In the small distances, i.e., intra-organization distances and up to the commute distance we do notice a shift in the probabilities to form paired collaborations that favors the earlier period. The most obvious departure between the curves makes a set of pairs that are 10km - 200km apart. This span of distances could be considered the commute distance and going up to the boundary of a region or a county [21,22].

**The law of gravity**

After establishing the existence of a gravitational law in our geographically-layered collaboration network, we wish to estimate how well does our data conform with this gravitational law. We linearize Newton's law of universal gravitation so we first write it log-transformed

$$(1) \quad log(F) = log(G) + log(M_i) + log(M_j) - 2log(r)$$

then, we perform a multivariate regression analysis to recover the gravitational constant G and estimate the correspondence of the other parameters, namely $M_i$, $M_j$ and $r$.

The masses $M_i$ and $M_j$ are the number of people holding faculty position in the respective interacting institutes $i$ and $j$. The distance $r$ is like in previous analyses of this paper. The regression analysis aggregates masses and distances by city location.

The equation we estimate is

$$(2) \quad log(F_{ij}) = \alpha + \beta_1 log(M_i) + \beta_2 log(M_j) + 2\gamma log(r)$$

and, in line with the theory, we hypothesize that $\beta_1 = \beta_2$ and $\gamma < 0$.

The regression results appear in Table 1. The negative impact of distance is well captured and so is the symmetrical influence of the main effects $M_i$, $M_j$ and $r$. The proportion constant is estimated as exp(0.72) = 2 and each main effect contributes to a $[((101/100)^{0.05}) - 1] * 100 = 0.05$ percent change in the force.

Continuing along this line of investigation we want to estimate whether there was a significant change in the equation's coefficients over time. In Table 2 we visualize side by side, two models that can be described by (2), each carries data from another time frame: "before" (<1995) and "after" (>=1995). The models give similar estimates to the main effects, however the intercept, or the "gravitational constant" is halved over time. It is noteworthy that the effects not described by the formula are essentially encapsulated into this constant. We could again imagine what is factored in: transportation and communication costs. Thus, before 1995 the geographical reach of collaboration was limited because it was harder to bridge distances, both long and short. To finalize, we estimated the sensitivity of $G$ to sub-periods outside of the internet era frame. The result indicates that $G$ does not change over periods preceding the mid-'90s. Thus, these results make $G$ likely to incorporate the environmental effects of the '90s as discussed above. (See Table 9 and supporting text in Methods)

Table 1: The gravitational law (1) estimated from our data using linear models of the main effects $M_i$, $M_j$ (the number of faculty on the flanks of a collaboration link), and the geographical distance across the link, $r$.

| | log(Force) | | |
|---|---|---|---|
| **Coefficient** | **Estimates** | **Conf. Int (95%)** | **P-Val** |
| Intercept | 0.72 | 0.52 – 0.92 | **<0.001** |
| log (1+Mi) | 0.05 | 0.02 – 0.07 | **0.001** |
| log (1+Mj) | 0.05 | 0.03 – 0.08 | **<0.001** |
| 2 log(r) | -0.06 | -0.06 – -0.05 | **<0.001** |
| Observations | 538 | | |
| $R^2$ / adjusted $R^2$ | 0.419 / 0.416 | | |

## 3. Discussion

A growing body of literature in complex networks has shown that social ties, be it in social Networks [13] mail exchanges [14] mobile phone communication [15,16] and blogs [15] follow a gravitational law. Namely, that the probability for a social tie between agents decays with a power of their distance. Focusing on scientific collaboration, Hoekman et al. [2] used data on publications between 313 regions in 33 European countries for the period 2000–2007 and found that the bias to collaborate with geographically close partners did not decrease, while the bias towards collaboration across territorial borders did decrease over that period. From the same time, Kumar et al. find

a similar pattern for scientific collaboration [1]. In this setting, our paper contributes to supporting a notable change in geographic collaboration as it is the first attempt to directly compare between the pre- and post-Internet eras.

Table 2: Similar to Table 1, two linear models were estimated for two time periods: before 1995 and after. The greatest difference between the models is the intercept (0.69 vs. 0.38). This difference is highlighted

| | log(Force) before 1995 | | | log(Force) after 1995 | | |
|---|---|---|---|---|---|---|
| **Coefficient** | **Estimates** | **Conf. Int (95%)** | **P-Val** | **Estimates** | **Conf. Int (95%)** | **P-Val** |
| Intercept | 0.69 | 0.45 – 0.93 | **<0.001** | 0.38 | 0.11 – 0.66 | **0.006** |
| log (1+Mi) | 0.04 | 0.00 – 0.07 | **0.029** | 0.05 | 0.00 – 0.09 | **0.042** |
| log (1+Mj) | 0.04 | 0.01 – 0.07 | **0.021** | 0.05 | 0.01 – 0.09 | **0.021** |
| 2 log(r) | -0.05 | -0.06 – -0.04 | **<0.001** | -0.04 | -0.05 – -0.03 | **<0.001** |
| Observations | 397 | | | 242 | | |
| $R^2$ / adjusted $R^2$ | 0.345 / 0.340 | | | 0.367 / 0.360 | | |

Several key attributes of the scientific production process suggest a strong tendency for geographical clustering. Clustering allows scientists to exploit spillovers and peer effects, which is in large part what motivated the emergence of scientific hubs such as the Silicon Valley, the Silicon Fen around Cambridge, England, or Tsukuba Science City in Japan. Moreover, scientific projects are social endeavors that share similar attributes to other, more tangible production processes. Hence, they rely on *division of labor*, entail repeating and ongoing interactions between collaborators, and require efficient management and monitoring. These ingredients are greatly facilitated with geographical clustering. Moreover, the ever-increasing competition for scientific recognition, the abundance of scientific fraud, and the rapid expansion of many scientific disciplines resulted in increased heterogeneity of core capabilities among scientists. Thus, issues of asymmetric information and distrust began to play role. In fact, the inability to establish trust within research teams was found to be among the most important factors that seal projects' fate to failure [16] In the Organization Science literature [17-19] it is well known that teams and organizations are plagued with policing, monitoring, and coordination costs, which are the result of the need to control and minimize these problems. Geographical proximity is therefore helpful in alleviating issues related to asymmetric information.

Against this backdrop, we explore a collaboration network at the individual level and support it with city-level analysis. The social, organizational and economic forces reviewed above promote geographical concentration of scientific collaboration. Next, we fit Newton's gravitational equation (1) onto the collaboration network using (2). The collaboration network's nodes, $i$ and $j$, are aggregated per city, the number of professors in a city is a proxy for the node's mass, $M_i$ and $M_j$, the distance between nodes, $r$, is the geographical distance, and the force, $F$, is the link strength. In this context a comparative estimation of two time periods is given: before and after the popularization of Internet use. We find that: (**i**) the goodness of fit is relatively high, (**ii**) the structural parameters, which measure the partial elasticity of collaboration strength with respect to distance and mass, remain stable over time, and (**iii**) the entire effect of the presumed environmental change brought over by the emergence of the Internet is bore by the intercept term, $G$.

Moreover, $G$, which is also the proportion of F over $M_iM_j/r^2$ decreases over time as distances between collaborators expand while not affecting the median geo-distance between nodes. We argue that this change corresponds to reduction in communication costs, including airfare, that followed the popularization of the Internet.

**Limitations and points to consider in the future**

In future research, we aim to generalize this study by investigating multi-author (>2) collaboration patterns, and adding information sets to tie further the connection between collaboration and communication costs.

Next, the current design cannot fully capture a cause and effect of the internet. Creating a time series analysis where the explanatory variable is internet use should be the design of choice. That design will provide insight in two main realms: determining (a) the causal effect of the introduction of the Internet, and (b) the cutoff year range.

Also, as we did not perform direct analysis of internet utilization among collaborating scientists, we believe that to a first approximation our results provide good estimation of this effect. Conducting a complementary analysis with proper variables might be useful to further support our findings. Last, since $G$ is a black box variable it encapsulates many latent effects. In a future study we will focus on decomposing this coefficient.

## 4. Methods

Our dataset is made of research articles and letters written in English, which appeared in the weekly edition of Nature over the period 1984 through 2015, enabling a direct comparison between two periods, immediately before and after the technological leap we collectively term "the Internet". Out of 3,207 articles and 24,121 letters published in this period we cleared away misclassification of author counts, missing authors' affiliations and geocoding problems to end up with 3,718 2-author titles, 6,626 authors and 2,426 uniquely identified addresses. The data retrieval procedure is summarized in Table 8. In our data, each title appears in context of a time span. For this reason, we extracted the affiliations of the authors particular to each paper, such that over time authors may not have accumulated old and redundant affiliation addresses. In cases where authors had more than one affiliation under a single title, we chose the first one.

Using the publicly available geocoding services of Google Maps, Bing, googlePlaces, and opencage we determined the geographic coordinates of each specific address. During the geocoding process, many queries returned no results, primarily for addresses in the former Soviet Union, old naming conventions like the Free Republic of Germany, and other countries where institutions have relocated since the 1990s, which preceded the Google era. The overall number of undirected links is 1,855,384, across which we computed distances on a great circle, using Napier rule. Thus, the natural weight for this collaboration network is the geodesic distance metric on the sphere. The frequency of papers in the 1990's has decreased drastically, while the frequency of letters is on a constant decline from 1000 in 1984 to about 700 in 2015. So, a cut in any year 1992-1999 will have generated roughly the same amount of papers and letters in the "before" and the "after" groups (Figure 4).

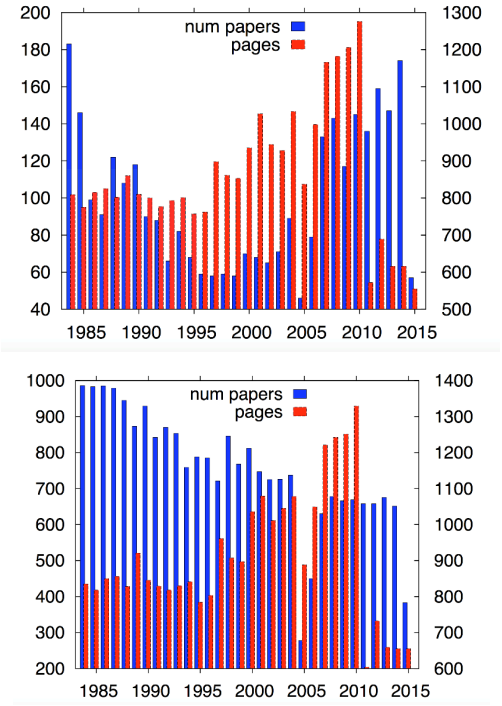


**Figure 4.** The number of *Nature* papers (top) and letters (bottom), including number of pages per year.

From this set of pairs, we generated distance distributions sliced by time periods 1984-1994, the "before", and 1995-2015, the "after" period. The probability density and cumulative density distributions (CDF) were subsequently retrieved and plotted. To test for stochastic dominance of pairs to occur across all distances we used Kolmogorov Smirnoff test between the two CDF curves. After analyzing the cumulative results, we aggregated the distances of pairs per city in the following manner:

author number 1 was assigned to the city of his/her affiliation. Then, for each city we estimated the mean and median distance to any of the second authors that collaborate with that city.

After establishing that the characteristic distance was 100km in both periods, with slight preference for smaller mean distance in the former period, we partitioned the distances between authors to "close" and "far", where "far" was determined as farther than the radius of the first author's city. The area of each city in the world was excavated from Wikipedia, and each city was modeled as a circle with radius $r = \sqrt{area/\pi}$.

To test whether any distance between collaborators and any of the time periods are dependent we used a Chi-square test of independence of these factors. We enumerated the pairs that collaborated in *Nature* who originated from any city. The statistical test of independence of the factors far/close and before/after yields rejection of the null, suggesting a tendency for far away pairs to occur more than expected in the later period (Table 3).

Last, to analyze the chance of a single city to attract collaborations across time we tested the difference between second author occurrences of "close" vs. "far" in the two time periods using Fisher's exact test. The data from which these results come, appear in the supplemental file localities_dist_pairs.xlsx. The shaded red and blue cells mark geographical areas that either expanded or contracted based on a 10-fold change in the mean collaboration distance.

Table 3: enumeration of factors, with marginals. The blue shading marks a value greater than would be expected by chance by 3 standard deviations. $\chi^2$= 15.371, df = 1, p-value = 8.834e-05

| All cities | after | before | **Sum** |
|---|---|---|---|
| Far | 323 | 575 | 898 |
| Close | 818 | 1999 | 2817 |
| **Sum** | 1141 | 2574 | 3715 |

Generally, the tendency of a paper to result from a collaboration within the same institute is high. More often if this is the sole paper that was published from that institute, and much greater than the frequency of papers written by two authors that are more than 200km apart. However, when breaking this to "before" vs. "after" this pattern changes in the "after" period. More examples from the supplemental file of centers that will have mixed tendencies, i.e., either expand or contract their reach to collaborators are: Davis, California, where the distance to collaborators grew from 327km (12 papers) to 1625km (8 papers), while Heidelberg, Germany, has contracted from 125km (17 papers) to zero (9 papers). Jerusalem, Israel has grown from zero (5 papers) to 2820km (5 papers). London grew from 502km (78 papers) to 1945km (30 papers) while Cambridge has contracted from 1250km (112 papers) to 284km (34 papers). New York City shrunk from 919km (64 papers) to 226km (40 papers). In Zurich the mean distance grew from zero (13 papers) to 709km (15 papers). Moscow grew from 560km (13 papers) to 3616km (4 papers). Overall there were more centers harboring an increase in distance in the second period. Although less accurate due to its linear nature, it still supports our original many-body clustering claim (cf. Figure 3).

Table 4: regions with the greatest change in geographical attraction (measured by the mean distance to collaborators) before 1995 and after, in the 95% confidence interval.

| **Region** | **Odds ratio** | **P-val** |
|---|---|---|
| Massachusetts | 0.894 | 0.0006 |
| Moscow | 15.242 | 0.0374 |
| Ontario | 4.201 | 0.0382 |
| Pennsylvania | 4.790 | 0.0391 |

To gain further insight into the role that geography plays in the two periods, we ran a Fisher's exact test for each AreaName that asks: "in city X what is the chance of preserving the paper number ratio 'before' over 'after' across distance?" Four areas stand out with divergent patterns in the two periods. These regions, listed in Table 4, we could fondly term the "evolving centers of knowledge".

The reason to run this test per region (a state in USA/Canada or a province in Europe) is for statistical power. Fisher's exact test allows us to consider zeros as non-structural in the test design. Still, there is lack of power in the test since many regions will have had one or no papers at all in one of the periods. Further data should be incorporated in order to broaden this test.

Then we designed a test for corroborating the gravitational law evidenced by other studies. For this

we added the quantities $M_i$ and $M_j$ to designate the masses in (1). These quantities were adapted from a list of degree-granting institutes in the USA (https://nces.ed.gov/ipeds/datacenter), excluding medical schools. The information is included in the supplementary file Data_6-9-2019---868.xls

Table 5: summary statistics of the gravitational formula (1) components for the years before 1995

| | F | $M_i$ | $M_j$ | r |
|---|---|---|---|---|
| **Min** | 1.000 | 0 | 0 | 0.0 |
| **1st Qu.** | 1.000 | 499 | 499 | 0.0 |
| **Median** | 1.000 | 1517 | 1498 | 146.6 |
| **Mean** | 3.557 | 2207 | 2102 | 1004.2 |
| **3rd Qu** | 2.000 | 2842 | 2707 | 1495.3 |
| **Max** | 110.000 | 16483 | 16483 | 7766.8 |

We restricted the dataset to collaboration pairs that reside in the US. We then aggregated the authors by the location of the closest known institute to increase the power of our statistical test. Thus, in a pair of nodes, Node *i* is roughly the city location of author *i* and Node *j* is the city location of author *j*. The distance, *r*, is the geographical distance between the nodes calculated on a great circle as before. The strength of each link is the number of collaboration pairs between the two nodes. This quantity is placed on the response side of the formulation, and termed Force.

A competing theory may claim that the change in the intercept is a product of the evolution of the system over time, not necessarily related of the logic of the suggested cutoff. In order to see whether this is a relevant claim we further partitioned the sub-period preceding the original cutoff into two: 1984-1989 and 1990-1995. We expected that the intercepts will be comparable and so they were. Table 9 lists the two regression models of the gravitational formula estimated for these two sub-periods. Although the estimators of masses are marginally significant, the differences are, overall, minute.

To further support our analysis, we include summary statistics of the main effects in the gravitational formula. Tables 5 and 6 give comparable quantities except the distances, *r*, that are slightly more skewed in the second period. The two periods therefore maintain similar prerequisites to the gravitational constant, *G*.

Table 6: summary statistics of the gravitational formula (1) components for the years after 1995

| | F | $M_i$ | $M_j$ | r |
|---|---|---|---|---|
| **Min** | 1.000 | 0 | 0 | 0.00 |
| **1st Qu.** | 1.000 | 526 | 514 | 0.00 |
| **Median** | 1.000 | 1507 | 1489 | 96.72 |
| **Mean** | 2.669 | 2218 | 1936 | 972.67 |
| **3rd Qu.** | 2.000 | 2832 | 2440 | 1514.95 |
| **Max.** | 43.000 | 16483 | 16483 | 8162.36 |

Table 7: network stylized facts of the full data from *Nature* (1984-2015) and the 2-author subsample of the same time frames. Blue (orange) highlight marks the lower (higher) value in a before-after pair

| | any coauthors | | | two coauthors | | |
|---|---|---|---|---|---|---|
| **Variable** | **Full** | **≤1995** | **>1995** | **Full** | **≤1995** | **>1995** |
| **Papers** | 25187 | 11106 | 14081 | 3717 | 2575 | 1142 |
| **Nodes** | 133300 | 38866 | 99773 | 6626 | 4624 | 2163 |
| **Links** | 2116798 | 150582 | 1966303 | 3717 | 2575 | 1142 |
| **Diameter** | 30 | 39 | 28 | 5 | 4 | 3 |
| **Assortativity** | 0.7174 | 0.8518 | 0.7096 | 0.1891 | 0.2504 | 0.2282 |
| **Path length** | 7.7457 | 12.1723 | 7.4116 | 1.2044 | 1.1041 | 1.2206 |
| **Clustering coef.** | 0.7879 | 0.8512 | 0.7921 | 0 | 0 | 0 |
| **Clique number** | 482 | 147 | 482 | 2 | 2 | 2 |
| **Modularity** | 0.8504 | 0.9377 | 0.8580 | 0.9994 | 0.9988 | 0.9995 |
| **Maximal cliques count** | 28706 | 10770 | 17696 | 3613 | 2490 | 1129 |
| **Components** | 6345 | 4819 | 3956 | 3013 | 2134 | 1034 |
| **Cliques over papers** | 0.0191 | 0.0132 | 0.0342 | 0.0005 | 0.0007 | 0.0006 |

Table 8: Summary statistics of research papers and letters published in Nature between 1984-2015 and written by two collaborators

| | Nature.com research papers | Nature.com letters |
|---|---|---|
| number of papers | 321 | 3397 |
| 1984-1994 | 219 | 2217 |
| 1995-2015 | 102 | 1179 |
| 2-author addresses | 860 | 8938 |
| 2-author addresses cleaned (permanent addresses or notes or correspondence addresses removed) | 804 | 8177 |
| 2-author papers' addresses where both authors have valid addresses | 642 | 6794 |
| number of papers where each author has a different address | 79 | 1039 |
| number of author pairs with different addresses | 158 | 2078 |
| same-address pairs | 242 | 2358 |
| unique addresses (from the 2-author papers where each author has a different address) | 155 | 1901 |
| mean distance [km] | | |
| 1984-1994 | 1096.7 | 893.7 |
| 1995-2015 | 646.6 | 1063.7 |
| median distance [km] | | |
| 1984-1994 | 0 | 0 |
| 1995-2015 | 0 | 0 |

# 5. References


[1] A.B. Smith, C.D. Jones, and E.F. Roberts, "Article Title", Journal, Publisher, Location, Date, pp. 1-10.

[2] Jones, C.D., A.B. Smith, and E.F. Roberts, Book Title, Publisher, Location, Date.

[1] Kumar, R., Kaski, K., & Fortunato, S. (2012). World citation and collaboration networks: uncovering the role of geography in science. *Nature Scientific Reports 2*:902, 1-7

[2] Hoekman, J., Frenken, K., & Tijssen, R. Research collaboration at a distance: Changing spatial patterns of scientific collaboration within Europe. (2010). *Research Policy* 39 (5): 662-673

[3] Chandra, A., Hajra, K., Das, P. & Sen, P. (2007). Modelling temporal and spatial features of collaboration network. *Int. J. Mod. Phys. C* 18: 1157–1172

[4] Georghiou, L. (1998). Global cooperation in research. *Res. Policy* 27: 611–626

[5] Rosenblat, T. S. & Mobius, M. M. (2004). Getting closer or drifting apart? *Q. J. Econ.,* 119: 971–1009

[6] Havemann, F., Heinz, M. & Kretschmer, H. (2006). Collaboration and distances between german immunological institutes–a trend analysis. *J. Biomed. Discov. Collab.,* 1, 6

[7] Agrawal, A. & Goldfarb, A. (2008). Restructuring research: Communication costs and the democratization of university innovation. *Am. Econ. Rev.* 98: 1578–90

[8] Krings, G., Calabrese, F., Ratti, C. & Blondel, V. D. (2009). Urban gravity: a model for inter-city telecommunication flows. *J. Stat. Mech.*, L07003

[9] *Brunson JC, Wang X, Laubenbacher RC.* (2017). Effects of research complexity and competition on the incidence and growth of coauthorship in biomedicine. *PLoS ONE*, 12(3): e0173444

[10] Cordero RJ, de León-Rodriguez CM, Alvarado-Torres JK, Rodriguez AR, Casadevall *A.* (2016). Life Science's Average Publishable Unit (APU) Has Increased over the Past Two Decades. *PLoS ONE*, 11(6): e0156983

[11] Levsky ME, Rosin A, Coon TP, Enslow WL, Miller MA. (2007). A descriptive analysis of authorship within medical journals, 1995-2005. *South Med J. Apr.* 100(4): 371-5

[12] Wagner, C.S., & Leydesdorff, L. (2005). Network structure, self-organization, and the growth of international collaboration in science. Research policy,34(10), 1608-1618.

13. Liben-Nowell, D., Novak, J., Kumar, R., Raghavan, P. & Tomkins. A. (2005). Geographic routing in social networks. *Proc. Natl. Acad. Sci. USA* 102: 11623–11628

[14] Levy M., Goldenberg, J. (2014). The gravitational law of social interaction. *Physica A* 393: 418-26

[15] Lambiotte, R. et al. (2008). Geographical dispersal of mobile communication networks. *Physica A* 387: 5317–5325

[16] Bennett, L. M., Gadlin, H. (2012). Collaboration and Team Science: From Theory to Practice. *J. Investig. Med.* 60(5): 758-775

[17] Coase, R. The Nature of the firm. (1937). *Economica* 4: 386–405

[18] Demange, G. (2004). On group stability in hierarchies and networks. *J. Polit. Econ. 112*: 754-778

[19] Guesnerie, R. & Oddou, C. (1988). Increasing returns to size and their limits. *The Scand. J. Econ.* 1: 259-273

[20] Noulas A, Scellato S, Lambiotte R, Pontil M, Mascolo C. (2012). A Tale of Many Cities: Universal Patterns in Human Urban Mobility. *PLoS ONE* 7(5): e37027

[21] Kelman, G. Manes, E., Bree, D., and Lamieri, M. (2017). Unusual Spatial Patterns of Industrial Firm Locations Uncover their Social Interactions. *Proceedings of the 50th Hawaii International Conference on System Sciences*, 9p. doi:10.24251/HICSS.2017.213.

[22] Kelman, G. Bree, D. S., Manes, E. Lamieri, M. Golo, N. and Solomon, S. (2015). Dissortative From the Outside, Assortative from the Inside: Social Structure and Behavior in the Industrial Trade Network. *Proceedings of the 48th Annual Hawaii International Conference on System Sciences*. pp1606-1615, doi: 10.1109/HICSS.2015.194

Table 9: Similar to Table 2, two linear models were estimated. This time for two pre-internet sub-periods: before 1989 and 1990-1994. No appreciable difference between the models, intercept included.

| | **log(Force) 1984-1989** | | | **log(Force) 1990-1994** | | |
|---|---|---|---|---|---|---|
| **Coefficient** | **Estimates** | **Conf. Int (95%)** | **P-Val** | **Estimates** | **Conf. Int (95%)** | **P-Val** |
| Intercept | 0.51 | 0.23 – 0.79 | **<0.001** | 0.48 | 0.21 – 0.75 | **<0.001** |
| log (1+Mi) | 0.04 | 0.00 – 0.09 | **0.033** | 0.03 | -0.01 – 0.07 | 0.133 |
| log (1+Mj) | 0.03 | -0.01 – 0.07 | 0.119 | 0.03 | -0.01 – 0.09 | 0.165 |
| 2 log(r) | -0.04 | -0.05 – -0.03 | **<0.001** | -0.04 | -0.05 – -0.03 | **<0.001** |
| Observations | 261 | | | 228 | | |
| $R^2$ / adjusted $R^2$ | 0.303 / 0.295 | | | 0.296 / 0.287 | | |