Journal of Statistical Mechanics: Theory and Experiment
An IOP and SISSA journal

The geography of scientific productivity: scaling in US computer science

Rui Carvalho and Michael Batty

Centre for Advanced Spatial Analysis, 1-19 Torrington Place, University College London, WC1E 6BT, UK
E-mail: rui.carvalho@ucl.ac.uk and mbatty@geog.ucl.ac.uk

Received 4 April 2006
Accepted 5 October 2006
Published 25 October 2006

Online at stacks.iop.org/JSTAT/2006/P10012
doi:10.1088/1742-5468/2006/10/P10012

Abstract. Here we extract the geographical addresses of authors in the Citeseer database of computer science papers. We show that the productivity of research centres in the United States follows a power-law regime, apart from the most productive centres for which we do not have enough data to reach definite conclusions. To investigate the spatial distribution of computer science research centres in the United States, we compute the two-point correlation function of the spatial point process and show that the observed power laws do not disappear even when we change the physical representation from geographical space to cartogram space. Our work suggests that the effect of physical location poses a challenge to ongoing efforts to develop realistic models of scientific productivity. We propose that the introduction of a fine scale geography may lead to more sophisticated indicators of scientific output.

Keywords: critical phenomena of socio-economic systems, scaling in socio-economic systems, stochastic processes, new applications of statistical mechanics

ArXiv ePrint: physics/0603242

©2006 IOP Publishing Ltd and SISSA

Contents

1. Introduction

In the last decade, the analysis of mankind's scientific endeavour has become a rapidly expanding interdisciplinary field. This has been mainly due to the advent of comprehensive online preprint servers and paper repositories, from which patterns of productivity and collaboration networks of individual scientists can be readily ascertained [1]. The vast amount of available data raises the hope that scientists and policy makers will soon be able to gain unprecedented insights into the location of research centres and their productivity. Indeed, little is known today about the influence that geographical location may have on 'invisible colleges'[1] (but see [3, 4]). Conversely, we are only just beginning to uncover how the historical growth of these 'invisible colleges' generates heterogeneities in the physical location of research centres and, therefore, of the scientists themselves.

Previous investigations of bibliometric data [5] by physicists have followed two main directions. On one hand, efforts have focused on characterizing the topological structure of collaboration networks [6]–[9]. On the other, researchers have used tools of statistical physics to gain insight into the growth dynamics of scientific outputs [10]–[12]. Despite this considerable progress, the relation of collaboration networks to the productivity of scientists depends on the still poorly understood fine geographical location of research centres.

Matia *et al* approached the challenge of characterizing institutional productivity by analysing 408 US institutes for the 11 year period 1991–2001 [12]. They observe a bimodal distribution and conjecture that this is indicative of a clustering effect of institutes of two different size classes [12].

The characterization of spatial structures at large geographical scales has a long tradition. In 1971, Glass and Tobler were the first to apply the radial distribution function (or two-point correlation function, as it is known in astrophysics [13]) to the study of cities on a part of the Spanish plateau [14, 16]. They choose a 40 mile square, homogeneous in town size and density, and apply concepts developed in the study of the statistical mechanics of equilibrium liquids. Although their analysis does not detect clustering, we

[1] An invisible college is a loose network of researchers who 'communicate with each other and transmit information across the whole field (...) to monitor the rapidly changing research "front".' [2, p 35].

J. Stat. Mech. (2006) P10012

would expect the two-point correlation function to reveal patterns of concentration and clustering in data whose population sizes vary over many orders of magnitude.

Recently, Yook *et al* showed that the nodes of the Internet are embedded on a fractal support driven by the fractal structure of the population worldwide [17]. This suggests that, in spatial networks with strong geographical constraints, the nodes may not be distributed randomly in space [18], but may be clustered as a function of population density. Further, Gastner and Newman presented an algorithm based on physical diffusion to draw density equalizing maps, or cartograms, in which the sizes of geographic regions appear in proportion to their population or some other property [19]. Cartograms give us a tool for probing into the dependence of one spatial variable (e.g. cancer occurrences) upon another (e.g. population). In particular, processes which are spatially clustered, but dependent on population densities, are expected to display random spatial distributions once the data are transformed by the cartogram [19, 20].

In order to bring the productivity of research centres and their spatial interaction patterns under a single roof, we follow an approach that is different from, but complementary to the ones presented above. Indeed, research centres are not homogeneously distributed in geographical space and it is likely that location will impact on their productivity and the structure of collaboration networks. However, to fully understand the role of location on the production of science and its networks, one must first characterize the underlying spatial processes, and this is the road we take here. We therefore investigate scientific productivity as a function of fine scale geographical location. Furthermore, to underpin these results, we characterize the spatial point process generated by the physical location of research centres.

To investigate the role of fine scale geography in the production of science, one needs to analyse a large data set. Traditional investigations of bibliometric data have been carried out by analysing databases like PubMed, arXiv.org or Thomson ISI. However, these databases suffer from drawbacks. Either the data contain only the address of the first (PubMed) or corresponding author (arXiv.org), or researchers are not uniquely associated with their addresses (Thomson ISI).

A more promising source of data is the Citeseer digital library, created in 1998 as a prototype of Autonomous Citation Indexing [21]. Citeseer locates computer science articles on the web in Postscript or PDF format and extracts citations from and to documents [22]. Citeseer has made its metadata available online [34] and the inclusion of an address and affiliation fields for each author allows a first rigorous analysis into the geography of a very large bibliometric database.

2. Spatial structure

We studied the Citeseer metadata, which contains 716 772 records, some of which are repeated and some of which have authors with empty address fields. We considered the $N = 379\,111$ (52.9%) unique papers for which citeseer identifies all authors and their respective addresses. Out of these N unique papers, we analysed the $M = 128\,348$ ($p_{\mathrm{US}} = 33.9\%$) papers which have one or more US authors. Interestingly, p_{US}, is in reasonable accordance with Thomson ISI global indicators, which state that between 1997 and 2001, the United States output 34.86% of the world's highly cited publications [23].

J. Stat. Mech. (2006) P10012

Table 1. Most productive ZIP codes and respective universities.

Rank	Zip	Fractional count	Institution
1	15213	2343.36	Carnegie Mellon University
2	02139	1891.18	MIT
3	94305	1512.12	Stanford University
4	94720	1496.76	University of California, Berkeley
5	20742	1144.70	University of Maryland, College Park

For each paper, we extracted the five-digit ZIP code from each author's address field and geocoded this ZIP into a *(latitude, longitude)* pair of coordinates [35]. We identified ZIP codes from the address field, by using regular expressions to match a five-digit code (plus the optional four digit code, which we ignored) preceded or followed by a US state (or its abbreviation) or the acronym *USA*. This will leave out addresses like *Roma 00185, Italy* or *Israel 84105*, but will also fail to locate the address *Physics Department, Northeastern University, Boston MA USA* as it lacks a ZIP code. We restricted the analysis to the 48 conterminous US states plus the District of Columbia.

We identified a total of 116 771 distinct authors with a US address. Out of these, 103 928 (89%) list a single ZIP code in their address, 10 579 (9.1%) belong to institutions located in two ZIP codes and 2264 (1.9%) are located in three or more institutions.

2.1. Productivity of research centres

To investigate the concept of scaling in publication output of academic research centres, we computed the probability distribution of total paper output per ZIP code. We note that ZIP codes were not aggregated. If two research centres belonging to the same institution have addresses with distinct ZIP codes, we considered them as distinct centres. This has the disadvantage of possibly counting more than one research centre per institution (instead of aggregating both to the same institution). However, Citeseer covers scientific articles in the field of computer science and it would be the exception that one institution would have several geographically separated computer science centres.

Our analysis identified 3393 different ZIP codes that matched the US census bureau tables. We implemented a version of fractional counting [5, 24] to compute the productivity of US research centres. For every paper, we parsed each author's address field and extracted the ZIP codes therein (there may be more than one ZIP, if the author belongs to more than one US institution). Each occurrence of a ZIP code in an address field of a paper increments the productivity of the research centre physically located at that ZIP code by $1/\phi$, where the normalization factor ϕ is computed as follows. For every address field in the paper being analysed, we made $\phi := \phi + 1$ if the address contains no ZIP codes (i.e. it is a non-US address), or $\phi := \phi + m$ if the address contains $m \geq 1$ ZIP codes (in which case that specific author will belong to m distinct US institutions).

Identifying research centres by ZIP code has the advantage of simplifying the data parsing algorithm, which is why we preferred this method to others based on aggregation by host institution. However, the method is an approximation, as it cannot distinguish between non-US addresses.

Table 1 displays the five most productive ZIP codes and their host institutions. Interestingly, the two most productive institutions, Carnegie Mellon University and MIT

J. Stat. Mech. (2006) P10012

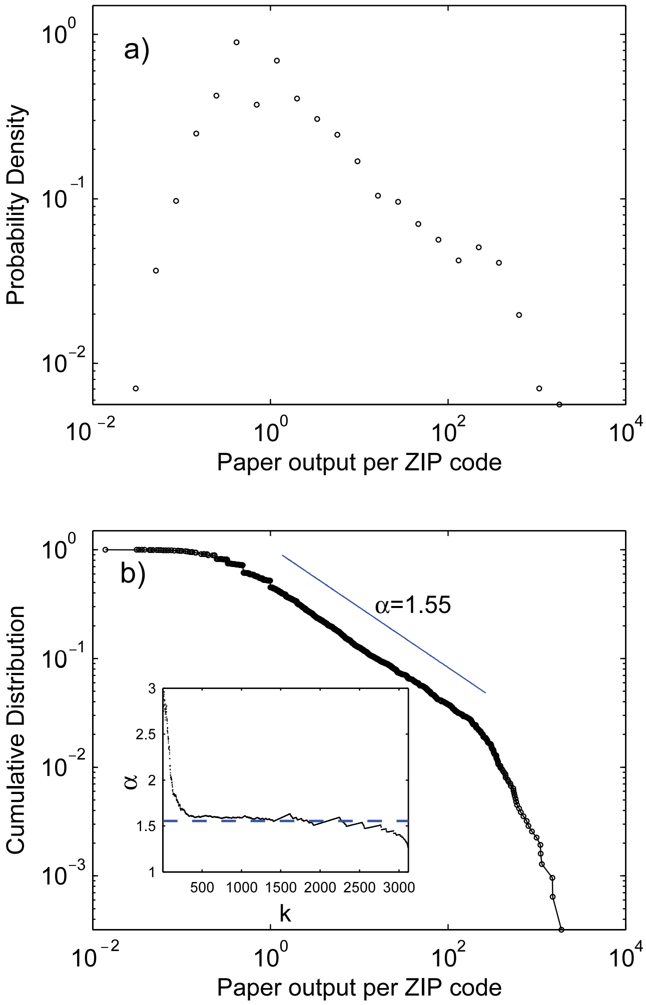

Figure 1. Probability distribution of paper output (fractional counts) per ZIP code. (a) The probability density is bimodal and can be approximated by a power-law regime between the two local maxima. (b) A least squares fit to the linear region of the cumulative distribution yields $\alpha = 1.55$. The inset shows the Hill plot [25] as the number of upper order statistics, k, is varied. The match between the plateau on the Hill plot and the least squares fit (dashed horizontal line) shows that our estimate of α is appropriate.

J. Stat. Mech. (2006) P10012

are also the two most acknowledged entities as shown by Giles and Councill in a previous study [21].

We then asked the question: what is the probability distribution of the research output of each research centre? To investigate this, we plot the probability density and cumulative distribution ($P[X > x] = \int_x^\infty p(y)\,\mathrm{d}y$) in figure 1. We found a bimodal probability distribution of research output by ZIP code (see figure 1(a)), in agreement with a previous study of the Thomson ISI database by Matia *et al* [12].

Our results suggest that this probability distribution displays power-law decay up to the 'knee' where the regime changes. The data were insufficient for determining whether the upper tail of the distribution also decays as a power law, albeit with a different exponent. This observation is in apparent contradiction with the findings of Matia *et al* who do not find a power-law regime. The authors examine the productivity of 408 US institutes, whereas our method revealed that papers had been output at 3393 US institutes. Therefore, the power-law decay which we observed may be due to our methodology which included all research institutes in the metadata. On the other hand, our analysis was limited in scope to the Citeseer database, whereas Matia *et al* analyse the Thomson ISI data set, hence comparisons with their wider study are necessarily inconclusive. Nevertheless, our results raise the question of whether power-law decay only appears once one is able to identify a large percentage of all research institutes.

2.2. The pulling power of research clusters

A simple point process in $\mathbb{R}^2$ may be considered as a random countable set $X \subset \mathbb{R}^2$. The first moment of a point process can be specified by a single number, the *intensity*, ρ, giving the expected number of points per unit area. The second moment can be specified by Ripley's K function [16], where $\rho K(r)$ is the expected number of points within distance r of an arbitrary point of the pattern.

The product density

$$\rho_2(\mathbf{x}_1, \mathbf{x}_2)\, \mathrm{d}A(\mathbf{x}_1)\, \mathrm{d}A(\mathbf{x}_2) = \rho^2 g(r)\, \mathrm{d}A(\mathbf{x}_1)\, \mathrm{d}A(\mathbf{x}_2) \tag{1}$$

describes the probability of finding a point in the area element $\mathrm{d}A(\mathbf{x}_1)$ and another point in $\mathrm{d}A(\mathbf{x}_2)$, at the distance $r = |\mathbf{x}_1 - \mathbf{x}_2|$, and $g(r)$ is the two-point correlation function. Ripley's K function is related to $g(r)$ by [26]

$$K(r) = 2\pi \int g(r)\, r\, \mathrm{d}r. \tag{2}$$

In other words, $g(r)$ is the density of $K(r)$ with respect to the radial measure $r\,\mathrm{d}r$. The benchmark of complete randomness is the spatial Poisson process, for which $g(r) = 1$ and $K(r) = \pi r^2$, the area of the search region for the points. Values larger than this indicate clustering on that distance scale, and smaller values indicate regularity.

The two-point correlation function can be estimated from N data points $\boldsymbol{x} \in D$ inside a sample window $\mathcal{W}$ by [27]:

$$g(r) = \frac{|\mathcal{W}|}{N(N-1)} \sum_{x \in D} \sum_{y \in D} \frac{\Phi_r(\boldsymbol{x}, \boldsymbol{y})}{2\pi r \Delta} \omega(\boldsymbol{x}, \boldsymbol{y}) \tag{3}$$

where $2\pi r \Delta$ is the area of the annulus centred at $\boldsymbol{x}$ with radius r and thickness Δ. Here $|\mathcal{W}|$ is the area of the sample window, and the sum is restricted to pairs of different points $\boldsymbol{x} \neq \boldsymbol{y}$. The function Φ_r is symmetric in its argument and $\Phi_r(\boldsymbol{x}, \boldsymbol{y}) = [r \leq d(\boldsymbol{x}, \boldsymbol{y}) \leq r + \Delta]$, where $d(\boldsymbol{x}, \boldsymbol{y})$ is the Euclidean distance between the two points and the condition in brackets equals 1 when true and 0 otherwise.

The function $\omega(\boldsymbol{x}, \boldsymbol{y})$ accounts for a bounded $\mathcal{W}$ by weighting points where the annulus intersects the edges of $\mathcal{W}$. There are a number of edge corrections available, but that of Ripley [15] has a long tradition both in human geography [16] and physics [27]:

$$\omega(\boldsymbol{x}, \boldsymbol{y}) = \frac{2\pi r}{F(\partial \mathcal{B}_r(\boldsymbol{x}) \cap \mathcal{W})} \tag{4}$$

J. Stat. Mech. (2006) P10012

where $\digamma(\partial\mathcal{B}_r(\boldsymbol{x}) \cap \mathcal{W})$ is the fraction of the perimeter of the circle $\mathcal{B}_r(\boldsymbol{x})$ with radius $r = |\boldsymbol{x} - \boldsymbol{y}|$ around $\boldsymbol{x}$ inside $\mathcal{W}$—e.g. $\digamma(\partial\mathcal{B}_r(\boldsymbol{x}) \cap \mathcal{W}) = \pi r$ if only half of the annulus falls inside $\mathcal{W}$. Note that $\omega(\boldsymbol{x}, \boldsymbol{y}) = 1$ iff $\partial\mathcal{B}_r(\boldsymbol{x}) \subset \mathcal{W}$, in which case the summand in (3) is simply the sum of $\Phi_r(\boldsymbol{x}, \boldsymbol{y})$ weighted by the area of the annulus centred at $\boldsymbol{x}$ with radius r and thickness Δ. If $\partial\mathcal{B}_r(\boldsymbol{x}) \cap \overline{\mathcal{W}} \neq \varnothing$, that is the circle $\mathcal{B}_r(\boldsymbol{x})$ is only partially in the sample window $\mathcal{W}$, then $\Phi_r(\boldsymbol{x}, \boldsymbol{y})$ is weighted by the area of the fraction of the annulus which is inside $\mathcal{W}$.

Of special physical interest is whether the two-point correlation is scale invariant. A scale-invariant $g(r)$ is an indicator of a fractal distribution of research centres, and is expected in critical phenomena [28].

To investigate the presence of power-law decay in the two-point correlation function we selected the 1046 research centres (ZIP codes) which had a total fractional count of two papers or more. We chose this productivity threshold for two main reasons. A first factor was to consider only research centres which can be clearly identified as active. Second, the computation of the two-point correlation function requires reasonable computer resources as $\mathcal{W}$ is a fine boundary of the United States—in our case, a polygon with 14 605 points.

Next, we projected the US map and the *(latitude, longitude)* pairs of the research centres with the Albers' equal area projection [29, 36] and computed the two-point correlation function, $g(r)$, of the resulting point process.

To investigate whether the decay of $g(r)$ is a function of the distribution of R&D funding or population, we applied several cartogram transformations to the base map and the points: first, we computed the cartogram projection using US R&D funding expenditure, by state, for the year 2001 [30, table B-17]; second we computed the cartogram with US population, by state and county, from the 2000 census [37]. The points representing the research centres were transformed accordingly to each cartogram. Figure 2(a) shows the Albers' equal area projection and each centre is represented by a circle with area proportional to the number of papers output on a logarithmic scale. Figures 2(b)–(d) show the cartograms with R&D expenditure by state, and population by state and county, respectively. It is obvious from these maps that as the cartogram transformation uses finer spatial scales (e.g. from US states to counties), the points become more homogeneously distributed spatially.

The two-point correlation function computed for the projected data (see figure 2(a)) is plotted in figure 3, where we observe a power-law decay $g(r) \sim r^{-\gamma}$ with $\gamma \simeq 1.16$. Next we asked the following question: can the power-law decay of $g(r)$ be explained by a clustering of research centres in areas where research funding or population is higher? To answer this question, we computed $g(r)$ for the same point process, but now using the data transformed by the cartograms with R&D expenditure by state (figure 2(b)), population by state (figure 2(c)), and population by county (figure 2(d)). Our results showed that the power-law decay was still present after the cartogram projections, although as the transformation was performed at finer spatial scales, $g(r)$ approached the expected value for a Poisson process, $g(r) = 1$, at shorter distances.

3. Discussion

Considerable advances have been made over the past few years in understanding the structure of scientific production and its networks. Along this road, physicists have

J. Stat. Mech. (2006) P10012

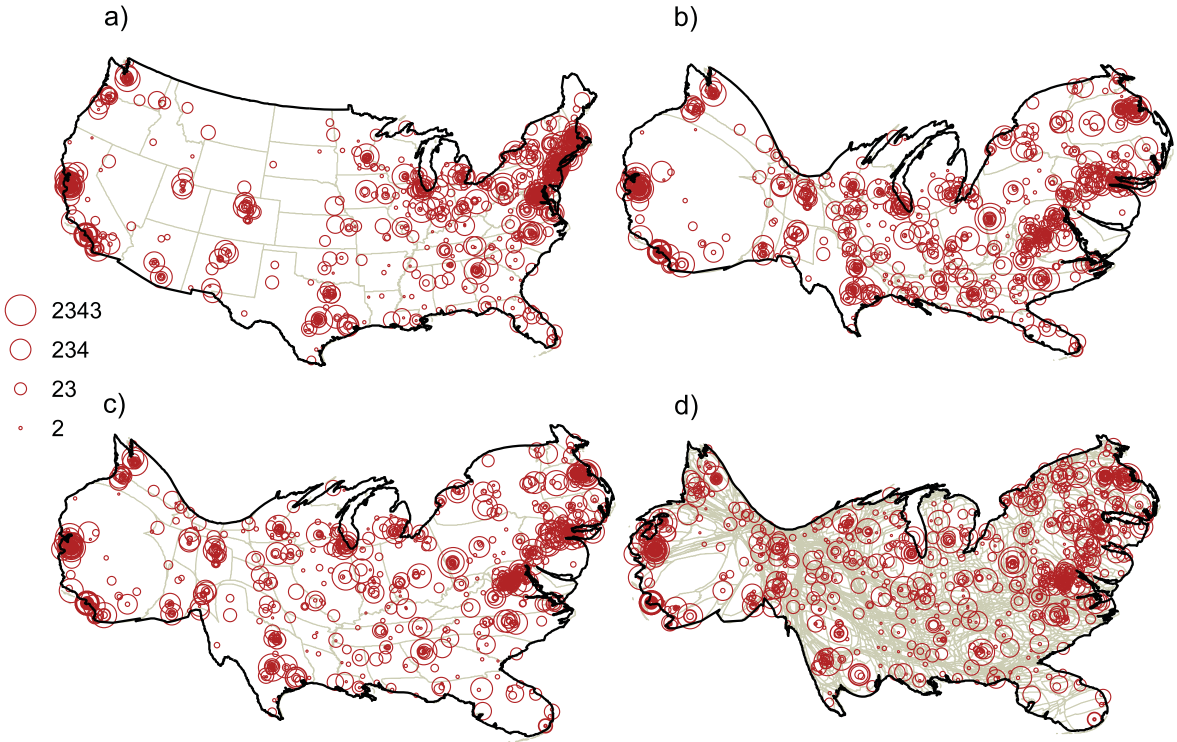

Figure 2. (a) Albers' equal area projection of the 48 conterminous states of the US plus the District of Columbia. Research centres are identified by circles with area proportional to their productivity on a logarithmic scale. ((b)–(d)) Data in (a) after a cartogram transformation with R&D expenditure by state (b), and population by state (c) and county (d), respectively. For each panel, we trace the 14 605-point border polygon used in the computation of the two-point correlation function.

computed a number of quantities for characterizing networks of scientific collaborations, mainly by analysing data from online preprint servers and repositories. However, these studies have not addressed the impact of fine scale physical location on the statistical characterization of the scientific enterprise and its networks. Here we have presented a detailed study of the productivity of research centres in US computer science (identified by ZIP codes) and characterized the pattern of spatial concentration which these centres display.

A first important conclusion of our study is that the productivity of US research centres in computer science was highly skewed. A surprising result of our study was the power-law decay of the probability distribution of research output for some orders of magnitude. A second important conclusion is that the physical location of research centres in the US formed a fractal set, which was not completely destroyed by population or research funding patterns.

Although we consider our results to be promising, there are still several caveats. First our conclusions are clearly only valid for the US [12, 31] and even from the Citeseer database, which we consider is the best currently available for such analysis, there are problems of missing and inaccurate data which we are not able to quantify. Nevertheless,

J. Stat. Mech. (2006) P10012

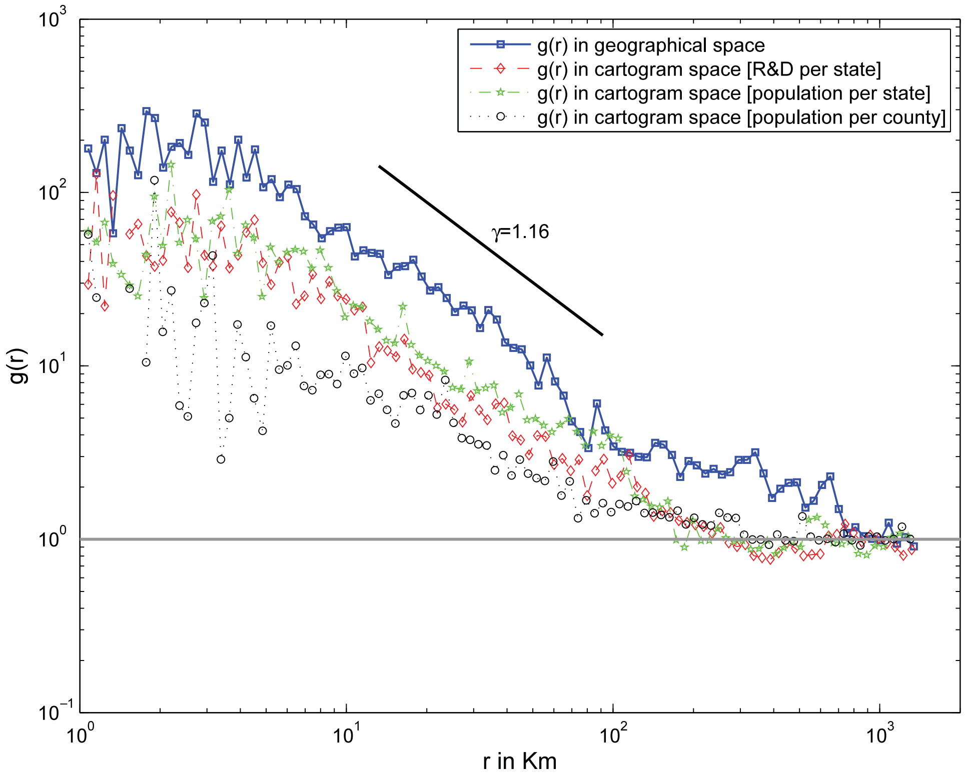

Figure 3. Variation of the two-point correlation function with distance (km). In blue, $g(r)$ computed from projections of the border and the points with the Albers' equal area projection. In red, green and black, $g(r)$ computed from further transforming the data by the cartogram projection with R&D per state, population per state and county, respectively. The horizontal line at $g = 1$ is the expected value of $g(r)$ for a Poisson process.

our results are consistent with those from the burgeoning geography of information technology which suggests in qualitative fashion, that such technologies are correlated with population but also have their own dynamic [32, 33]. In this sense, our result that the scaling inherent in the geographical distribution of paper production in US computer science is still present once the geography has been normalized with respect to the distribution of population and R&D expenditures, implies processes that are endogenous to the dynamics of research [11].

In summary, the method introduced in this paper could serve as a starting point for an investigation of the role of the fine scale physical location of research centres in the production of science. Our study focused on US computer science but further analyses should be possible as preprint server repositories make more elaborate metadata available. And such developments may lead to a better understanding of the role of physical location not just in science, but for a much wider class of complex spatial systems.

Acknowledgments

We wish to thank Michael Gastner (SFI) for help with the code used to generate cartograms and Isaac Councill (Penn State) for help with the Citeseer database. This

J. Stat. Mech. (2006) P10012

research was supported by the Engineering and Physical Sciences Research Council under grant EP/C513703/1.

References

[1] Shiffrin R M and Börner K, *Mapping knowledge domains*, 2004 *Proc. Nat. Acad. Sci.* **101** (Suppl. 1) 5183

[2] Crane D, 1972 *Invisible Colleges: Diffusion of Knowledge in Scientific Communication* (Chicago, IL: University of Chicago Press)

[3] Batty M, *The geography of scientific citation*, 2003 *Environ. Plan.* A **35** 761–5

[4] Börner K, Penumarthy S, Meiss M and Ke W, *Mapping the diffusion of information among major US research institutions*, 2006 *Scientometrics* **68** 415

[5] Egghe L and Rousseau R, *Introduction to informetrics: quantitative methods in library, documentation and information*, 1990 *Science* (Amsterdam: Elsevier)

[6] Newman M E J, *The structure of scientific collaboration networks*, 2001 *Proc. Nat. Acad. Sci.* **98** 404

[7] Newman M E J, *Scientific collaboration networks: I. Network construction and fundamental results*, 2001 *Phys. Rev.* E **64** 016131

[8] Newman M E J, *Scientific collaboration networks: II. Shortest paths, weighted networks, and centrality*, 2001 *Phys. Rev.* E **64** 016132

[9] Barabási A L, Jeong H, Néda Z, Ravasz E, Schubert A and Vicsek T, *Evolution of the social network of scientific collaborations*, 2002 *Physica* A **311** 590

[10] Plerou V, Amaral L A N, Gopikrishnan P, Meyer M and Stanley H E, *Similarities between the growth dynamics of university research and of competitive economic activities*, 1999 *Nature* **400** 433

[11] Amaral L A N, Gopikrishnan P, Matia K, Plerou V and Stanley H E, *Application of statistical physics methods and concepts to the study of science and technology systems*, 2001 *Scientometrics* **51** 9–36

[12] Matia K, Amaral L A N, Luwel M, Moed H F and Stanley H E, *Scaling phenomena in the growth dynamics of scientific output*, 2005 *J. Am. Soc. Inf. Sci. Technol.* **56** 893–902

[13] Peacock J, 1999 *Cosmological Physics* (Cambridge: Cambridge University Press)

[14] Glass L and Tobler W R, *Uniform distribution of objects in a homeogeneous field: cities on a plain*, 1971 *Nature* **233** 67

[15] Ripley B D, *The second-order analysis of stationary point processes*, 1976 *J. Appl. Probab.* **13** 255

[16] Ripley B D, *Modelling spatial patterns*, 1977 *J. R. Stat. Soc.* B **39** 172

[17] Yook S-H, Jeong H and Barabási A-L, *Modeling the internet's large-scale topology*, 2002 *Proc. Nat. Acad. Sci.* **99** 13382

[18] Boccaletti S, Latora V, Moreno Y, Chavez M and Hwang D-U, *Complex networks: structure and dynamics*, 2006 *Phys. Rep.* **424** 175

[19] Gastner M T and Newman M E J, *Diffusion-based method for producing density-equalizing maps*, 2004 *Proc. Nat. Acad. Sci.* **101** 7499

[20] Gastner M T and Newman M E J, *Optimal design of spatial distribution networks*, 2006 *Phys. Rev.* E **74** 016117 [cond-mat/0603278]

[21] Giles C L and Councill I G, *Who gets acknowledged: Measuring scientific contributions through automatic acknowledgment indexing*, 2004 *Proc. Nat. Acad. Sci.* **101** 17599

[22] Goodrum A A, McCain K W, Lawrence S and Giles C L, *Scholarly publishing in the internet age: a citation analysis of computer Science literature*, 2001 *Inf. Proc. Manag.* **37** 661

[23] King D A, *The scientific impact of nations*, 2004 *Nature* **430** 311

[24] de Solla Price D, *Letter to the editor*, 1981 *Science* **212** 987

[25] Dress H, Haan L d and Resnick S, *How to make a Hill plot*, 2000 *Ann. Stat.* **28** 254

[26] Stoyan D, *Basic ideas of spatial statistics*, 2000 *Statistical Physics and Spatial Statistics* ed K Mecke and D Stoyan (Heidelberg: Springer) p 3

[27] Kerscher M, Szapudi I and Szalay A S, *A comparison of estimators for the two-point correlation function*, 2000 *Astrophys. J.* **535** L13

[28] Kerscher M, *Statistical analysis of large-scale structure in the universe*, 2000 *Statistical Physics and Spatial Statistics* ed K Mecke and D Stoyan (Heidelberg: Springer) p 36

[29] Robinson A H, Morrison J L, Muehrcke P C, Kimerling A J and Guptill S C, 1995 *Elements of Cartography* (New York: Wiley)

[30] National Science Foundation, Division of Science Resources Statistics, *National Patterns of Research and Development Resources: 2003,* NSF 05-308, Brandon Shackelford (Arlington, VA 2005)

[31] Moed H F and Luwel M, *The business of research*, 1999 *Nature* **400** 411

J. Stat. Mech. (2006) P10012

[32] Zook M A, 2005 *The Geography Of the Internet Industry: Venture Capital, Dot-Coms, And Local Knowledge* (Oxford: Blackwell)
[33] Dodge M, *Understanding cyberspace cartographies: A critical analysis of Internet network infrastructure maps*, 2006 *PhD Thesis* University College London
[34] http://citeseer.ist.psu.edu/oai.html Accessed 22/02/2006
[35] http://www.census.gov/geo/www/tiger/zip1999.html
[36] http://www.census.gov/geo/www/cob/
[37] http://www.census.gov/popest/datasets.html

J. Stat. Mech. (2006) P10012