Evolutionary Events in a Mathematical Sciences Research Collaboration Network

J.C. Brunson[a,1], S. Fassino[b,1,3], A. McInnes[c,1,2,3], M. Narayan[d,1,3], B. Richardson[c,1,2,3], C. Franck[e,a], P. Ion[f], R. Laubenbacher[a,*]

[a] *Virginia Bioinformatics Institute, Washington St, MC 0477, Virginia Tech, Blacksburg, Virginia*
[b] *Department of Mathematics, 227 Ayres Hall, University of Tennessee, Knoxville, TN 37996*
[c] *Department of Mathematics and Computer Science, Oakwood University, Cooper Complex Bld. B, 7000 Adventist Blvd, Huntsville, AL 35896*
[d] *Lyman Briggs College, Michigan State University, 35 East Holmes Hall, East Lansing, MI 48825*
[e] *Laboratory for Interdisciplinary Statistical Analysis, 212 Hutcheson Hall (0439), Blacksburg, VA 24061*
[f] *Mathematical Reviews, P.O. Box 8604, Ann Arbor, MI 48107*

Abstract

Collaboration is key to scientific research, and increasingly to mathematics. This paper contains a longitudinal investigation of mathematics collaboration and publishing using the proprietary database *Mathematical Reviews*, maintained by the American Mathematical Society. The database contains publications by several hundred thousand researchers over 25 years. Mathematical scientists became more interconnected, collaborative, and interdisciplinary over this interval, and twice the network experienced dramatic structural shifts. These events are examined and possible external factors are discussed. Smaller subject-specific subnetworks exhibit behavior that provides insight into the aggregate dynamics. The data are available upon request to the Executive Director of the AMS.

Keywords: mathematics research, collaboration networks, evolving networks
2010 MSC: 91D30

1. Introduction

Collaboration networks have been studied extensively in recent years, thanks to the availability of several excellent databases, e.g. [Gro02, New01b, BJN+02, FLC+04, AOL+07, TL07, Per10]. These studies have revealed a diversity of topological structure, especially across disciplines, depicting typical ranges of basic graph-theoretic metrics across real-world networks. While longitudinal studies are increasingly common, they predominantly take a cumulative approach; they observe network growth after a designated starting year, which may then be compared to evolving graph models [BJN+02]. However, evolving real-world networks can take decades to exhibit clear long-term trends [RB10], and short-term changes in structure and behavior become obscured by aggregating information [TL07]. To strengthen models of scientific research collaboration, cumulative models must be supplemented by dynamic models that capture the *effective* relationships among researchers [TL07, KW06]. Furthermore, while collaboration networks are often treated in the larger context of complex networks, important differences exist between social networks and other real-world networks [NP03]. Most available publishing databases are too specialized (by discipline or region) to exhibit clear long-term trends.

A specialized theory of evolving social networks is therefore required, and is underway. In this paper we examine a large, longitudinal collabo-

*Corresponding author: Virginia Bioinformatics Institute, Washington St, MC 0477, Blacksburg, VA 24060. Tel.: (540) 231-7506. Fax: (540) 231-2606.
Email addresses: jabrunso@vbi.vt.edu (J.C. Brunson), sfassino@utk.edu (S. Fassino), antonio.mcinnes@oakwood.edu (A. McInnes), nadaraj2@msu.edu (M. Narayan), brya_08@yahoo.com (B. Richardson), chfranck@vbi.vt.edu (C. Franck), ion@ams.org (P. Ion), reinhard@vbi.vt.edu (R. Laubenbacher)

[1]These authors were partially funded by NSF Award:477855.
[2]These authors were partially funded by HHMI:52006309.
[3]These authors contributed equally to this project.

[1]These authors were partially funded by NSF Award:477855.
[2]These authors were partially funded by HHMI:52006309.
[3]These authors contributed equally to this project.

*To whom correspondence should be addressed. Address: Bioinformatics Facility, Washington Street, Postal Code 0477; phone: (540) 231-7506; fax: (540) 231-2606; email: reinhard@vbi.vt.edu.

ration network. The American Mathematical Society (AMS) maintains the proprietary database *Mathematical Reviews* (*MR*), and we study this database across 1985–2009, during which nearly 430,000 authors produced nearly 1.6 million publications. *MR* aims to catalogue every mathematical sciences publication each year, including both print and online journals, books, proceedings, and other publications [Jac97]. We therefore treat the database as a census of the literature; however, we caution that the mathematics literature is itself a fraction of the broader scientific literature and highly entangled therewith.[4] The network is much larger than most studied scientific collaboration networks, extends over a longer time, and is of consistently great size, which will allow us to characterize long-term trends and fluctuations.

2. Materials and Methods

Our data consist, for each publication, of encoded author IDs, subject classifications from the AMS Mathematics Subject Classification Scheme [Soc11], and the year of publication. While authors and publications, taken together, exhibit a bipartite structure, and bipartite models that preserve this structure show promise [BMG03, GMY05, ZWL$^+$08], the larger literature and better-understood statistical toolkit on unipartite models allows us to better contextualize our network. We therefore adopt a unipartite model. In this model, nodes $v_1, \ldots, v_n$ correspond to authors and links $v_i v_j$ (m total) indicate coauthorship. Each link $v_i v_j = v_j v_i$ receives a (collaboration) weight w_{ij} given by the number of joint publications by v_i and v_j [New04]. The graph evolves over time as authors begin and cease publishing.

We investigated the evolving topology of the network using several well-understood graph-theoretic metrics. To account for the publishing process underlying this structure while maintaining our unipartite perspective, we introduced publication-sensitive analogs to the strictly graph-theoretic assortativity and clustering coefficients. These metrics reveal network properties not captured by the originals and may warrant further use.

[4] Our network not necessarily more bibliographically complete or self-contained than previously studied collaboration networks (such as the Los Alamos preprint archive in [New01b]); non-mathematician authors may appear on mathematics publications no less frequently than physicists who abstain from online databases collaborate with authors who do not.

Table 1: The *MR* network over two intervals.

MR network	1940–2000	1985–2009
years	61	25
papers	1598	1599
authors	337	429
avg. authors/paper	1.45	1.75
avg. papers/author	6.9	6.5
collab. pairs	496	876
avg. no. coauthors	2.9	4.1
prop. in largest comp.	.62	.75
avg. separation	7.56	7.31
global clustering coeff.	.15	.14
avg. clustering coeff.	.34	.61
assortativity	.12	.069

Subject classifications within the *MR* database include two-digit prefixes from 01 to 97. We divided the literature coarsely into "pure" (03–58) and "applied" (60–95) subnetworks and for some specific analyses into the similarly-sized subclassifications indicated in Fig. 2.[5]

To trace the effective structure of these networks, we used, depending on the metric, nonoverlapping intervals of one year or of five years or sliding windows of 5 years. The choice of 5-year intervals offers meaningful comparisons to [New01b]. In plots, we identify each window by its last year; for instance, the year 1997 may refer to the interval 1993–7. Because the network grows most quickly from 1985 to 1989, and because data is not complete in the most recent years, we focused mainly on the period 1989–2007. The smaller subnetworks fluctuated widely, obscuring long-term trends, but their behavior illuminates trends in the aggregate by distinguishing the disciplines most reflective of, and plausibly responsible for, those trends.

3. Trends in Mathematical Publishing

We examined long-term trends exhibited by the *MR* network. We present the publishing data in a raw statistical analysis, emphasizing the relationship of output rates to coauthorship and to multidisciplinarity.

Table 1 compares our network (all 25 years taken together) with the *MR* network studied in [Gro02]. Several differences detectible in the table reflect long-term trends discussed below, including increased collaboration (rows 4, 6, 7) and greater network connectivity (rows 8, 9, 11).

[5] This scheme is imperfect. For instance, much of 60 (Probability Theory and Stochastic Processes) might be classified as pure mathematics, but this would split 60 from 62 (Statistics).

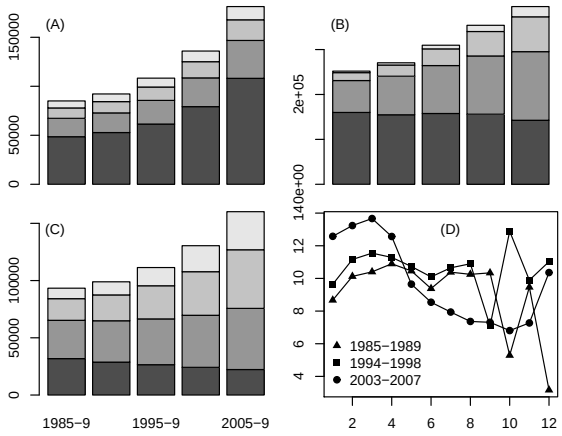

Figure 1: Across nonoverlapping 5-year intervals: (A) Numbers of authors with 1, 2, 3, and 4 publications. (B) Numbers of publications by 1, 2, 3, and 4 authors. (C) Numbers of authors with 0, 1, 2, and 3 coauthors. (D) Average number of publications by authors of a paper, as a function of the number of authors on the paper.

3.1. Publishing rates

We measured publishing rates individually and collaboratively. Mathematics researchers have grown more numerous and collaborative at accelerating rates, though without becoming steadily more prolific (Fig. 1 (A–C)). In fact, in recent years highly collaborative projects have involved authors less prolific within mathematics, and average prolificity has declined (Fig. 1 (D) and 3 (A–C)). While the *number* of more prolific authors has accelerated, it has been outpaced by the number of authors of only one publication, as we discuss in the supplementary text. These trends were starker in the applied network, which housed a greater proportion of less prolific authors, reversed its trend from more to less prolific years earlier than the pure, and a greater surge in one-time authors (Fig. 3 (A–F)). Credit for declining average publishing rates therefore rests largely with such authors.

This surge in less prolific authors reflects a major event around 2001 that we will describe further. A closer look reveals another event years earlier: a surge in collaborative publishing after 1995. From the interval 1989–95 to the interval 1995–2009, rates of 2- to 6-author publications rose and rates of 7- and more-author publications reversed from decline to rise (Fig. 2 (G)). Fluctuations in subject classification assignments and in graph-theoretic structure illuminated these events, as we discuss in the next section.

3.2. Multidisciplinarity

The literature grew steadily more multidisciplinary, except for a brief period of specialization

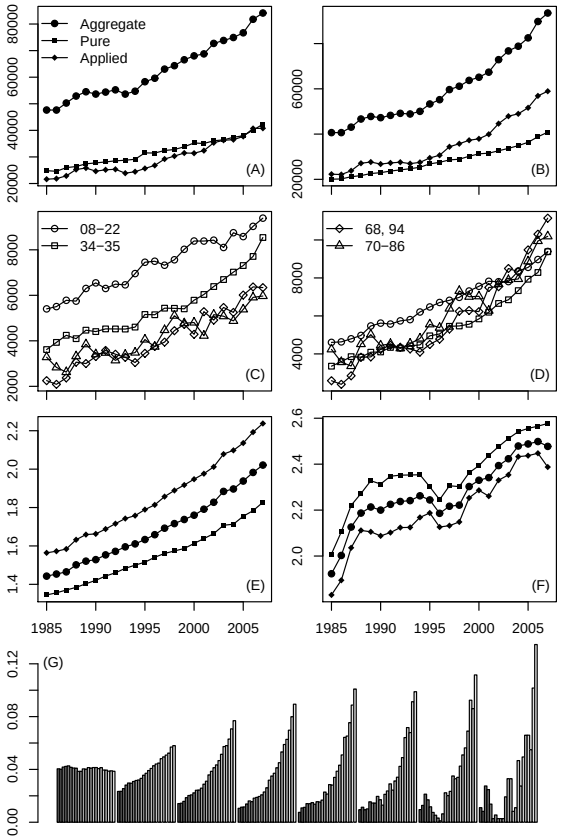

Figure 2: Across 1-year intervals, 1985–2007: (A) Number of publications. (B) Number of authors. (C) Number of publications across subnetworks. (D) Number of authors across subnetworks. (E) Average number of authors per publication. (F) Average number of subject classifications per publication. (G) For each fixed number of authors (1–8), a histogram over years of publications attributed to that number of authors. Each histogram is scaled by the total number of publications by that number of authors. We include four subdisciplines: algebra (○; 08–22), differential equations (□; 34–35), computer science and information (◇; 68, 94), and classical physics (△; 70–86).

near 1995, as measured by the average number of secondary subject classifications $\langle s \rangle = \langle s_i \rangle_i$ (as i ranges across publications). Meanwhile the average number of secondary authors per publication $\langle a \rangle = \langle a_i \rangle_i$ (authors beyond the requisite one) increased monotonically (Fig. 2 (E,F)). While the applied network exhibited larger $\langle a \rangle$ but smaller $\langle s \rangle$, a regression model reveals a *positive* relationship between s_i and a_i that is stronger in the more multidisciplinary pure network (Fig. 3 (G–I)). We fit to the combined pure and applied literature the linear model

$$s_i = \alpha_0 + \alpha_1 a_i + \alpha_2 u_i + \alpha_3 a_i u_i + \epsilon_i, \qquad (1)$$

where the indicator u_i takes the value 0 if the publication is classified as pure and 1 otherwise. The parameter α_1 is then the effect of a_i in the pure network, $\alpha_1 + \alpha_3$ that of a_i in the ap-

plied, and α_3 the interaction effect of a_i and u_i. This coauthorship–multidisciplinarity relationship weakened over time, but the subnetworks grew variably similar and dissimilar over different intervals. Shifts in α_3 coincided with the two events: the pure and applied networks grew similar after the earlier event but dissimilar after the second.

4. Evolution of the Coauthorship Graph

We adopt a graph-theoretic approach to study connectivity, correlations, and clustering in terms of coauthorship. We made use of several graph-theoretic metrics. We performed calculations on largest connected components unless otherwise noted, for two reasons: (1) The same fluctuations are visible in time series for entire (disconnected) graphs, though often subdued. (2) The steadily shrinking proportion of nodes outside largest components affects statistics sensitive to the presence of isolated authors and to highly connected, independent teams (two common forms that small connected components take).

4.1. Individual and network connectivity

We measured connectivity three principal ways. The number k_i of coauthors of an author v_i is that author's *degree*, a measure of individual connectivity. With an increase in average degree comes an increase in graph *density* $D = m/\frac{1}{2}n(n-1)$, the proportion of possible node–node links that are realized. We may also measure global connectivity by the proportion of nodes subsumed by the largest connected component itself. Finally, we gain insight into the efficiency of this connectivity from the mean node–node separation within this component. We adopted the *harmonic mean separation* $\langle \ell \rangle$ between pairs of authors defined by

$$\langle \ell \rangle^{-1} = \sum_{i,j} {\ell_{ij}}^{-1} / \tfrac{1}{2}n(n-1),$$

rather than the arithmetic mean, to place emphasis on local connections [LM01]. (The metrics are nonetheless highly correlated. Taking their residuals from linear fits each year as ordered pairs produces $r = .995$, though the arithmetic mean varies more about its fit.)

While the average degree $\langle k \rangle = 2m/n$ increased, the rate of increase over 1994–2009 was almost double that over 1989–1994, predominantly due to applied publications (Fig. 1 (C) and 3 (D,E)). The change in pace of average degree after 1994, especially in the applied network, is consistent with the surge in collaboration observed above. Meanwhile, the largest component of the aggregate network absorbed greater proportions of authors, from 37% (1989) to 65% (2009), though this trend decelerated. These proportions span the typical range for collaboration networks [Gro02, New01b, BJN$^+$02, TL07, Per10], suggesting that the proportional rise will continue to decelerate as the networks approach a practical upper limit on collaboration network cohesion. The pure and applied subnetworks conglomerated similarly, though they exhibited different mean separation, with the applied network consistently more dispersed (Fig. 3 (F)). Generally, $\langle \ell \rangle$ decreases as D increases, and while residuals from linear fits of these metrics exhibited some correlation ($r = -.63$), the pure network was consistently tighter despite the greater density of the applied. This implies that the structure of collaboration varies in important ways, among disciplines and over time, and we studied this structure through coauthor correlations and clustering.

4.2. Correlations among collaborators

A network is *assortative*, or exhibits *assortative mixing*, when similar pairs of nodes are preferentially linked, disassortative when linking is preferentially dissimilar, and nonassortative otherwise [New03]. The normalized degree correlation coefficient r_{col} measures assortative mixing by number of collaborators. We supplemented r_{col} with a measure r_{pub} of assortative mixing by number of publications. (See the supporting information for a formal definition.)

Collaboration networks are known to be assortative by collaborators ($.1 < r_{\text{col}} < .4$) but previous studies indicate that mathematics networks are less so [TL07, New03]. We also found r_{col} to be positive but low in the aggregate, pure, and applied networks, though some subdisciplines were largely nonassortative (Fig. 4 (A,D)). Mathematics researchers were more strongly correlated by publishing rate ($.3 < r_{\text{pub}} < .6$). The applied network and subnetworks exhibited stronger correlations by both metrics, signifying more hierarchical organization.

The events come into sharper focus through these correlation coefficients. Around 1995 the network shifted from progressively disassortative mixing to progressively assortative, mostly with respect to collaborators and predominantly among applied researchers. After 2001 this trend reversed again as coauthors became less correlated with respect both to collaborators and to publications, and in the latter case earlier in the pure network.

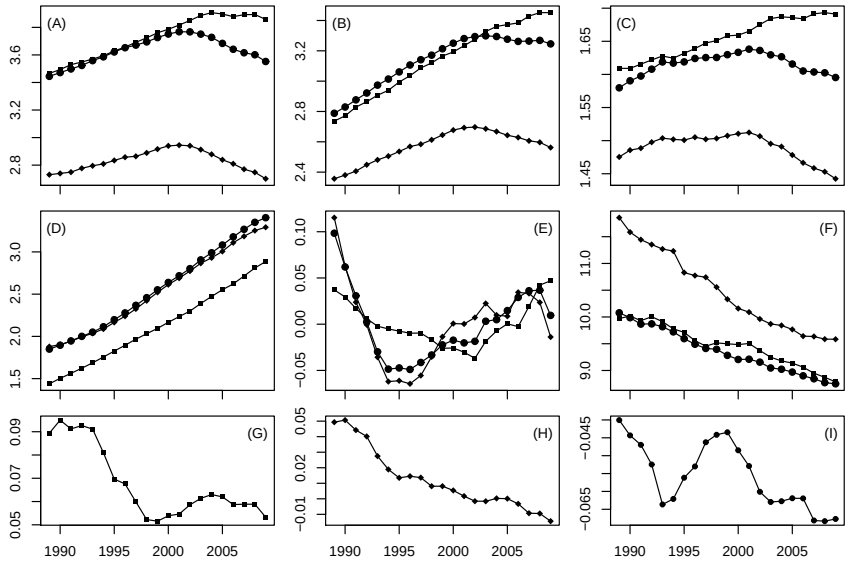

Figure 3: Across 5-year sliding windows, 1985–9 to 2005–9: Average number of (A) publications per author, (B) collaborative publications per author, and (C) publications per pair (zeros omitted from each average). (D) Average degree $\langle k \rangle$. (E) Residuals of $\langle k \rangle$ from best linear fits. (F) Harmonic average separation. (G–I) Estimates of α_1, $\alpha_1 + \alpha_3$, and α_3 in the regression model (1).

4.3. Scale-freeness

Recently the graph-theoretic statistic $s(g) = \sum k_i k_j$, a sum taken over edges of graph g, has been used to quantify "scale-freeness" among graphs with a common (scaling) degree sequence [LADW05]. This s-metric is greatest where high-degree nodes are linked preferentially, producing a highly connected "hub-like" core. The metric

$$S(g) = \frac{s(g) - s_{\min}}{s_{\max} - s_{\min}}$$

(refined in [Li07]) normalizes s over the range of s-values across graphs of the same degree sequence as g, and therefore has range $[0, 1]$. S may also be interpreted as a similar normalization of r_{col} across this collection of graphs.

The S-metric is best understood across graphs with a power law degree sequence, and while the degree sequences of collaboration networks are not well-modeled by power laws [LADW05, New01c], power-law approximations are popular [Gro02, New01b] and helpful in distinguishing collaboration networks from other categories of networks [ASBS00]. Recent studies apply S to several model networks [LADW05, THL06, THLH07, BC08] but applications to social networks are limited [Hsi09]. We observed $.48 < S < .58$. The time series for S reveals that fluctuations in r_{col} may be interpreted in the context of gradually diminishing scale-freeness (Fig 4 (C)).

4.4. Clustering

Whereas $\langle k \rangle$, r_{col}, r_{pub}, and S measure individual and pairwise structure, clustering measures structure concerning triples. Among triples of authors a, b, and c where a and b collaborated and a and c collaborated, the *(global) clustering coefficient* C expresses the proportion for whom b and c also collaborated. Disassortative graphs permit a reduced number of possible triangles among nodes of different degrees, and thus admit a smaller range of values for C. This may be globally accounted for using relative probabilities [New01a], which we consider in the supplementary text, but we principally adopted a clustering coefficient $\tilde{C}$ designed to correct for this locally [SV05]. The aggregate network showed low clustering, $.22 < C < .27$, compared to other coauthorship graphs [New01c, BJN+02, TL07, Per10]. The correction doubled the range to $.46 < \tilde{C} < .53$, as observed in other networks [SV05].

We also introduced an *exclusive clustering coefficient*: Among triples of authors where a and b collaborated without c, a and c collaborated without b, and both b and c published at least twice, $C_\times$ is the proportion for whom b and c collaborated without a. $C_\times$ detects changes in coauthorship that cannot be explained by team collaboration, and distinct pairwise publications suggest stronger, transitive relationships than single common publications.[6]

Locally, the clustering coefficient c_i of an author v_i is the proportion of the $k_i(k_i - 1)/2$ pairs

[6] While clustering coefficients have been introduced for bipartite author–publication graphs [RA04, ZWL+08], they do not address this issue directly.

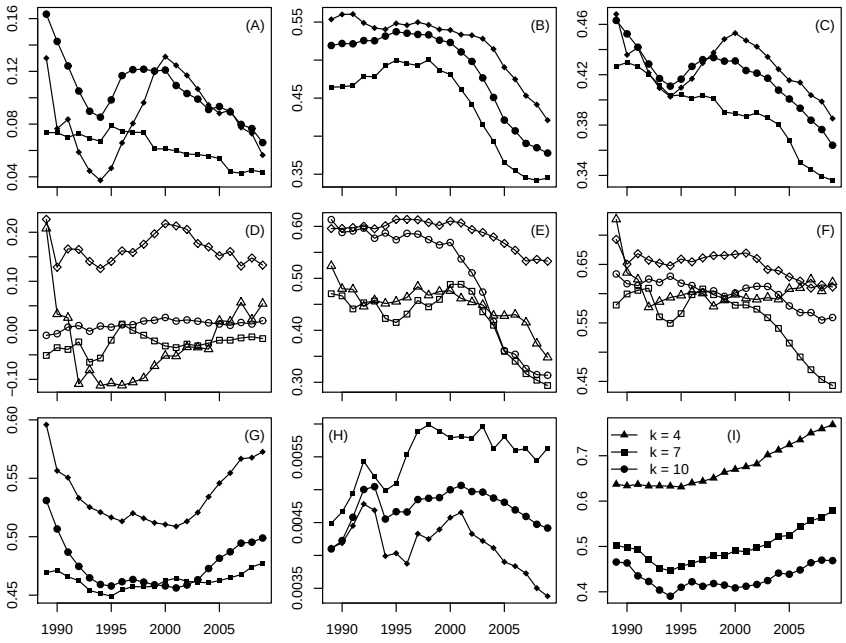

Figure 4: Across 5-year sliding windows, 1985–9 to 2005–9: (A) Assortative mixing by number of coauthors, r_{col}. (B) Assortative mixing by number of publications, r_{pub}. (C) The S-metric. (D) r_{col} across subdisciplines. (E) r_{pub} across subdisciplines. (F) S across subdisciplines. (G) Global clustering $\tilde{C}$ corrected for degree assortativity: (H) Global clustering $C_\times$ based on pairwise exclusive publications. (I) Average local clustering corrected for degree assortativity across authors of fixed degree 4, 7, and 10.

of their collaborators who have themselves collaborated. Again we adopted the correction $\tilde{c}_i$ from [SV05]. We used the network-wide average $\langle \tilde{c} \rangle$ in our change point analysis (Table 2), and we stratified authors by degree in Fig. 4 (I) to compare clustering across differently-connected researchers.

We found patterns of clustering to reaffirm that the two events were driven by larger teams of collaborators. While $C_\times \leq C \leq \tilde{C}$ by definition, in our network C_X was comparatively tiny, with $.0041 < C_\times < .0051$ (Fig. 4 (G,H)). This indicates that highly collaborative projects drove overall clustering behavior. Clustering increased after 1995, at both local and global scales and by both graph-theoretic and exclusive definitions. In particular, better-connected authors exhibited greater clustering earlier than less-connected authors. After 2001, however, graph-theoretic clustering surged while exclusive clustering plummeted (Fig. 4 (G–I)). After 2001 the pure and applied networks grew increasingly dissimilar, with greater graph-theoretic clustering in the applied but greater exclusive clustering in the pure. This suggests a connection to the more prominent disassortative mixing in the applied network, and indeed the propagation of highly collaborative projects by disassortative short-lived research teams would explain both dissimilarities. Furthermore, $C_\times$ mimicked collaboration weight $\langle w \rangle$ and r_{pub}, suggesting that autonomous collaborations are better forged among similarly prolific authors.

5. Events and Change Point Models

While long-term trends varied widely, fluctuations in our metrics, as revealed by residuals from linear fits, were often highly correlated (Fig. 5). Similar fluctuations suggest mathematical or sociological dependencies among properties; we grouped together metrics with strongly correlated time series and identify these groups by symbol in Table 2 and Fig. 6 (C,D). We used a *change point model*[7] to arrange these shifts chronologically.

Our change point model fits a continuous, piecewise-linear curve with one corner to a set of ordered pairs (x_i, y_i), subject to error from a fixed distribution, analogously to a linear fit. The model takes the slopes, intercept, and change point to be unknown and the errors to come from a normal distribution with unknown variance:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 (x_i - c)\delta_{x_i > c} + \epsilon_i, \;\; \epsilon_i \sim N(0, \sigma)$$

[7] An recent bibliography of change point problems by Khodadadi and Asgharian [KA08] traces change point models to a 1954 discussion by Page [Pag54] on piecewise continuous models.

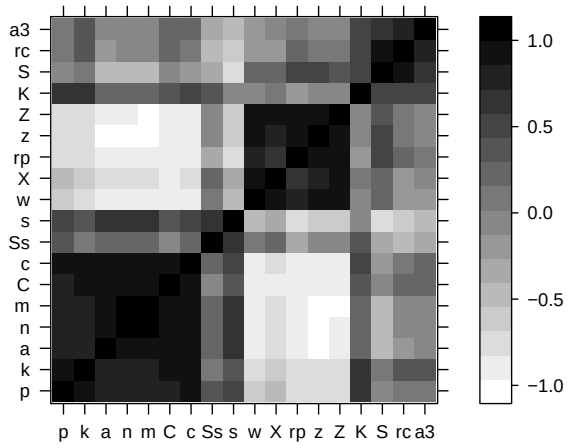

Figure 5: Correlation matrix of time series of key statistics across 5-year sliding windows. Top/right to bottom/left: p, no. publications; k, $\langle k \rangle$; a, $\langle a \rangle$; n, n; m, m; C, $\tilde{C}$; c, $\langle \tilde{c} \rangle$; Ss, avg. no. subject classifications per author; s, $\langle s \rangle$; w, $\langle w \rangle$; X, $C_\times$; rp, r_{pub}; z, avg. no. publications per author; Z, avg. no. collab. publications per coauthor; K, $\langle \kappa \rangle$; S, S; rc, r_{col}; a3, α_3.

The indicator $\delta_{x>c}$ takes the value 1 when $x > c$ and 0 otherwise. The parameters $\beta_0, \beta_1, \beta_2, c$ encode the two slopes β_1 and $\beta_1+\beta_2$, the y-intercept β_0, and the change point c. Our code in R uses iterative methods to find estimators for the parameters that minimize $SSE = \sum_i {\epsilon_i}^2$. We optimized this model over intervals visually centered about the dramatic shift of each time series to obtain the dates in Table 2. We used intervals of 11 years when possible, shortening to 10 years in case our algorithm failed, for consistency with sliding windows. We exhibit code and all change point fits to time series in the supplementary materials.

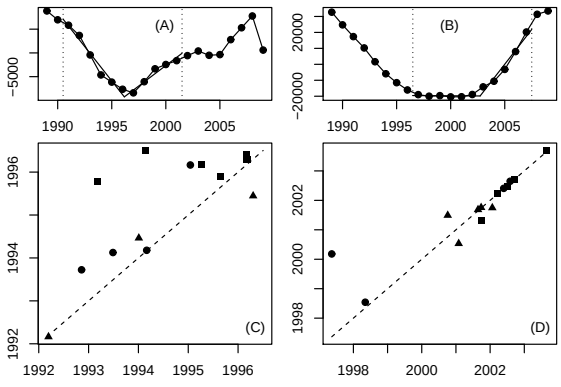

Figure 6: Across 5-year sliding windows, 1985–9 to 2005–9: (A) Residuals from best linear fit of the number of publications, with change point fits about the mid-90s event overlaid. (B) Residuals from best linear fit of the number of authors, with change point fit about the mid-2000s event overlaid. (C–D) Differences in best-fit change points from the aggregate network to the few-author network at the both events.

Table 2: Change points in time series around both events.

	statistic (symbol groups as in Fig. 6)	change
▲	avg. collab. publications/coauthor	1992.13
•	correlation by no. coauthors (r_{col})	1993.13
■	global SV-clustering coeff. ($\tilde{C}$)	1993.17
•	scale-freeness (S)	1993.49
■	avg. no. coauthors ($\langle k \rangle$)	1994.13
•	diff. in collab.-multidisc. effects (α_3)	1994.17
▲	avg. publications/author	1994.30
•	avg. no. collaborations ($\langle \kappa \rangle$)	1995.04
■	avg. SV-clustering coeff. ($\langle \tilde{c} \rangle$)	1995.26
■	avg. authors/publication ($\langle a \rangle$)	1995.65
■	no. publications	1996.16
■	no. collab. pairs (m)	1996.17
■	no. authors (n)	1996.19
▲	correlation by no. publications (r_{pub})	1996.30
•	correlation by no. coauthors (r_{col})	1997.37
•	diff. in collab.-multidisc. effects (α_3)	1998.35
▲	correlation by no. publications (r_{pub})	2000.76
▲	global exclusive clustering coeff. ($C_\times$)	2001.08
▲	avg. collab. publications/coauthor	2001.65
■	global SV-clustering coeff. ($\tilde{C}$)	2001.74
▲	avg. publications/author	2001.74
▲	avg. collab. weight ($\langle w \rangle$)	2002.06
■	avg. SV-clustering coeff. ($\langle \tilde{c} \rangle$)	2002.22
•	scale-freeness (S)	2002.40
■	avg. authors/publication ($\langle a \rangle$)	2002.51
•	avg. no. collaborations ($\langle \kappa \rangle$)	2002.59
■	no. collab. pairs (m)	2002.72
■	no. authors (n)	2003.65

We also contrasted the aggregate network with a "few-author" network constructed from publications of 6 authors or fewer, which would be unaffected by the reversals of trends exhibited in Fig. 2 (C). For uniformity in our change point analysis we drew all statistics from largest components, so time series for many statistics differ from those presented earlier. (In Table 2, the *multidegree* of a node v_i in a weighted graph is the sum κ_i of the weights of its links; we then say that author v_i has engaged in κ_i "collaborations".) The few-author network exhibited fluctuations similar to, but not always simultaneous with, those of the aggregate. By several metrics it experienced the first event later than the aggregate but the second event at essentially the same time (Fig. 6 (C,D)). This suggests that highly collaborative projects were inceptive to the first event while not necessarily to the second.

The chronologies of both events, as arranged in Table 2, suggest "top-down" narratives, with shifts in hierarchical metrics sensitive to highly central or prolific authors preceding shifts in metrics of local connectivity, and shifts in network-wide averages and totals manifesting last.

6. Discussion

Over 25 years the mathematics collaboration network grew steadily larger, more collaborative, and better-connected both locally and globally. While the applied network was better connected locally ($\langle a \rangle$, $\langle k \rangle$, $\langle \tilde{c} \rangle$) and exhibited more hierarchical structure (r_{col}, r_{pub}, S), the pure network was better connected globally ($\langle \ell \rangle$) and exhibited stronger local connections ($\langle w \rangle$, $\langle s \rangle$, α_3). In particular, while the small-world properties of low mean separation and high clustering have been reproduced together by a variety of real-world and model networks [ASBS00] [WS98, Jac08, BM10], neither of our major subnetworks is clearly the superior "small world" of the two.

The mid-90s event was characterized by proliferated and strengthened collaboration (Fig. 2 (E), $\langle k \rangle$, Fig. 4 (G,H)), a weakening relationship between collaboration and multidisciplinarity (Fig. 3 (G–I), and moderately increased assortative mixing (Fig. 4 (A,C)). The rise in several-author publications explains the stabilization of clustering; exclusive clustering had already been rising (Fig. 4 (G,H)). However, increases in clustering and hierarchical metrics were still evident in the network constructed from 6- or fewer-author publications. The delay in shifts from the aggregate to this few-author network indicates that highly collaborative projects were inceptive to the event (Fig. 6 (C)), a proposition supported by the "top-down" progression of change ponits.

These qualities of the event, the similar behavior of the pure and applied disciplines, and timing suggest a possible factor: the rise of e-communications and the World Wide Web. Among academic Internet milestones are the introductions of the `arXiv` in 1991, which went online in 1993 [Gin09], and of `MathSciNet` in 1996, which made the *MR* publishing database available through a graphical web interface [Jac97]. We should expect researchers in more applied subdisciplines, who historically made greater use of computing resources, to have made quicker use of these tools, and indeed the applied network and its subdisciplines exhibited the above trends more clearly (Fig. 4 (A–F)).

The early-2000s event tells a dissimilar story. This event was characterized by weakening average publishing rates and collaboration strength (Fig. 3 (A–C) and 4 (H)) due in part to an influx of less prolific authors (Fig. 1 (D) and 3 (D)) and dramatically disassortative mixing (Fig. 4 (A–C)). While disassortativity was ubiquitous, lower publishing rates were more evident in applied disciplines. Increased clustering was largely explained by a further acceleration in several-author publications (Fig. 4 (G,H)). Highly collaborative projects were not so inceptive (Fig. 6 (D)).

The growth in the research community and interconnections within it, simultaneous with weakening average publishing rates and collaboration rates, may be largely explained by the surge in transient authors. This surge may reflect an increasing trend toward interdisciplinary research involving many researchers outside mathematics who publish seldom but in larger teams. This is consistent with the absence of a specialization trend during this event, which distinguishes it from the earlier event (Fig. 2 (F)). A possible contributing factor to such a trend would have been an increased emphasis on interdisciplinary projects at funding agencies such the National Science Foundation, the largest funder of U.S. mathematics research. We note that the event was concurrent with increased funding by the NSF for its Division of Mathematical Sciences [nsf11] recommended by a 1998 report [Odo98]. (See the supplementary text for detailed discussion.) A change point fit to 5-year funding averages places the surge at 2001.33, toward the beginning of the event (Table 2). NSF funding affects almost exclusively U.S.-based research, however, while the *MR* database covers worldwide output.

7. Conclusions

The community of researchers in mathematical sciences has grown at an increasing rate since 1985, and their research output has accelerated. Amidst this growth the literature has become increasingly multidisciplinary and the network of researchers has grown better-connected and individual researchers more collaborative. Increased collaboration has been due in large part to highly collaborative teams of researchers, many of whose members have short mathematical publishing histories. Such disassortative authorship has been more prevalent in applied disciplines, which nonetheless exhibit more hierarchical organization, while researchers in more pure disciplines maintain longer collaborations and are less separated by degrees of coauthorship. The network drastically reorganized twice between 1985 and 2009, in different ways that suggest dissimilar causes and consequences.

The *MR* network is huge and admits much more analysis than we have performed. Data collected since 1940 are being processed and will be released soon, which will allow investigators to treat the database from conception. We omitted discussion of linking mechanisms, and of a range of tools for detecting community structure, for which the

MR network holds great potential. We suggested possible partial explanations for the two events we described, but it is beyond the scope of this paper to consider these hypotheses thoroughly. More detailed information on mathematical publishing and its funding may be obtained from the *MR* database and from government agencies, and follow-up investigations may provide deeper insights into these possible connections.

Acknowledgments. The authors are grateful to the American Mathematical Society for providing access to the *MR* database and for agreeing to make the data publicly available (by request to the Executive Director). The authors thank Sastry Pantula and Philippe Tondeur for providing information on NSF funding for mathematics, and Sid Redner, Ritchie C. Vaughan, Betsy Williams, and fellow participants of the Summer 2010 REU in Modeling and Simulation in Systems Biology for helpful conversations and support.

References

[AOL+07] Juan A. Almendral, J.G. Oliveira, L. López, J.F.F. Mendes, and Miguel A.F. Sanjuán. The network of scientific collaborations within the european framework programme. *Physica A: Statistical Mechanics and its Applications*, 384(2):675–683, 2007.

[ASBS00] L. A. N. Amaral, A. Scala, M. Barthélémy, and H. E. Stanley. Classes of small-world networks. *PNAS*, 97(21), 2000.

[BC08] Isabel Beichl and Brian Cloteaux. Measuring the effectiveness of the s-metric to produce better network models. In *Proceedings of the 40th Conference on Winter Simulation*, WSC '08, pages 1020–1028. Winter Simulation Conference, 2008.

[BJN+02] A. L. Barabási, H. Jeong, Z. Néda, E. Ravasz, A. Schubert, and T. Vicsek. Evolution of the social network of scientific collaborations. *Phys. A*, 311(3-4):590–614, 2002.

[BM10] Dionysios Barmpoutis and Richard M. Murray. Networks with the smallest average distance and the largest average clustering, 2010.

[BMG03] Katy Börner, Jeegar T. Maru, and Robert L. Goldstone. The simultaneous evolution of author and paper networks, November 2003.

[FLC+04] Ying Fan, Menghui Li, Jiawei Chen, Liang Gao, Zengru Di, and Jinshan Wu. Network of econophysicists: a weighted network to investigate the development of econophysics. *International Journal of Modern Physics B*, 18(17–19):2505–2511, 2004.

[Gin09] Paul Ginsparg. The global village pioneers. *Learned Publishing*, 22(2):95–100, 2009.

[GMY05] Michel L. Goldstein, Steven A. Morris, and Gary G. Yen. Group-based yule model for bipartite author-paper networks. *Phys. Rev. E*, 71(2):026108, Feb 2005.

[Gro02] J. W. Grossman. The evolution of the mathematical research collaboration graph. *Congressus Numerantium*, pages 201–212, 2002.

[Hsi09] Ping-Nan Hsiao. A social network model based on topology vision. In *Complex (2)*, pages 1398–1409, 2009.

[Jac97] Allyn Jackson. Chinese acrobatics, an old-time brewery, and the "much needed gap": the life of Mathematical Reviews. *Notices Amer. Math. Soc.*, 44(3):330–337, 1997.

[Jac08] Matthew O. Jackson. Average distance, diameter, and clustering in social networks with homophily. In *Proceedings of the 4th International Workshop on Internet and Network Economics*, WINE '08, pages 4–11, Berlin, Heidelberg, 2008. Springer-Verlag.

[KA08] A. Khodadadi and M. Asgharian. Change-point problems and regression: An annotated bibliography. *Collection of Biostatistics Research Archive*, 2008.

[KW06] Gueorgi Kossinets and Duncan J. Watts. Empirical analysis of an evolving social network. *Science*, 311(5757):88–90, 2006.

[LADW05] Lun Li, David Alderson, John C. Doyle, and Walter Willinger. Towards a theory of scale-free graphs: definition, properties, and implications. *Internet Math.*, 2(4):431–523, 2005.

[Li07] Lun Li. *Topologies of complex networks: functions and structures.* PhD thesis, California Institute of Technology, 2007.

[LM01] Vito Latora and Massimo Marchiori. Efficient behavior of small-world networks. *Phys. Rev. Lett.*, 87(19):198701, Oct 2001.

[New01a] M. E. J. Newman. Clustering and preferential attachment in growing networks. *Phys. Rev. E*, 64:025102, 2001.

[New01b] M. E. J. Newman. Scientific collaboration networks. i. network construction and fundamental results. *Phys. Rev. E*, 64(016131), 2001.

[New01c] M. E. J. Newman. The structure of scientific collaboration networks. *Proc. Natl. Acad. Sci. USA*, 98(2):404–409 (electronic), 2001.

[New03] M. E. J. Newman. Mixing patterns in networks. *Phys. Rev. E (3)*, 67(2):026126, 13, 2003.

[New04] Mark E. J. Newman. Who is the best connected scientist? A study of scientific coauthorship networks. In *Complex networks*, volume 650 of *Lecture Notes in Phys.*, pages 337–370. Springer, Berlin, 2004.

[NP03] M. E. J. Newman and Juyong Park. Why social networks are different from other types of networks. *Phys. Rev. E*, 68(036122 [8 pages]), 2003.

[nsf11] Nsf budget requests to congress and annual appropriations, August 2011.

[Odo98] William E. Odom. Report of the senior assessment panel for the international assessment of the u.s. mathematical sciences. Technical Report NSF 98-95, National Science Foundation, March 1998.

[Pag54] E. S. Page. Continuous inspection schemes. *Biometrika*, 41(1/2):pp. 100–115, 1954.

[Per10] M. Perc. Growth and structure of slovenias scientific collaboration network. *Journal of Informetrics*, 4(4):475–482, 2010.

[RA04] Garry Robins and Malcolm Alexander. Small worlds among interlocking directors: Network structure and distance in bipartite graphs. *CMOT*, 10(1):69–94, 2004.

[RB10] Martin Rosvall and Carl T. Bergstrom. Mapping Change in Large Networks. *PLoS ONE*, 5(1):e8694+, January 2010.

[Soc11] American Mathematical Society. 2010 mathe-

matics subject classification, June 2011.

[SV05] Sara Nadiv Soffer and Alexei Vázquez. Network clustering coefficient without degree-correlation biases. *Phys. Rev. E*, 71(5):057101, May 2005.

[THL06] Yihjia Tsai, Ping-Nan Hsiao, and Ching-Chang Lin. A social network model based on caveman network. In *Communications and Networking in China, 2006. ChinaCom '06. First International Conference on*, pages 1 – 6, oct. 2006.

[THLH07] Yihjia Tsai, Ping-Nan Hsiao, Ching-Chang Lin, and Wen-Fa Huang. Node degree sequence preserving and controlling s-metric. In *Proceedings of the 2007 IEEE ICACT*, pages 2157–2161, 2007.

[TL07] Marco Tomassini and Leslie Luthi. Empirical analysis of the evolution of a scientific collaboration network. *Physica A: Statistical Mechanics and its Applications*, 385(2):750–764, 2007.

[WS98] D J Watts and S H Strogatz. Collective dynamics of small-world networks. *Nature*, 393(6684):440–442, 1998.

[ZWL+08] Peng Zhang, Jinliang Wang, Xiaojia Li, Menghui Li, Zengru Di, and Ying Fan. Clustering coefficient and community structure of bipartite networks. *Physica A: Statistical Mechanics and its Applications*, 387(27):6869–6875, 2008.

8. Supporting Information

We performed calculations in `R`, including graph-theoretic calculations using the `igraph` package and some original code. All code is available upon request from the first author.

8.1. Publishing rates and connectivity

As discussed in the main text, one-time authors increasingly worked in large research teams (Fig. S1 (A)). The also comprised an increasing proportion of the community after 2000 (Fig. S1 (D)), while the proportion consisting of more prolific authors (3 or more publications) decreased (Fig. S1 (C)).

Network density (Fig. S1 (B)) is directly related to average degree. The relevance of the pure–applied split is evident in the higher density in both, indicating stronger connectivity among pure researchers and applied researchers separately than among all mathematics researchers. The fluctuations, especially in the applied network, were similar to those in r_{col}. The aggregate, pure, and applied networks exhibited similar cohesion with respect to largest connected components (Fig. S1 (C)), and the "S"-shape of the curves (especially that of the applied) suggests that this proportion is approaching its practical limit.

8.2. Assortative mixing by publications

Newman [New03] defines a normalized degree correlation coefficient by way of "remaining degree": Starting with a pair of nodes (v_i, v_j), take the number of neighbors of each excluding the other, $(k_i - 1, k_j - 1)$. These are their remaining degrees. We define r_{pub} analogously using the notion of *remaining prolificity*.

Consider authors v_i and v_j who have authored z_i and z_j publications, respectively, and have collaborated on w_{ij} of them. z_i is then the "prolificity" of v_i (and z_j that of v_j) while w_{ij} is the "collaboration weight" of v_i and v_j together. Define the *remaining prolificity* of v_i with respect to v_j to be $z_i - w_{ij}$, the number of publications by v_i not coauthored with v_j. Since in graph-theoretic language we say that v_i is adjacent to the link (v_i, v_j), we refer to the remaining prolificity of the *adjacency* of node v_i to link (v_i, v_j).

Where the network includes n_x authors of prolificity x, set $p_x = n_x / \sum_{x'} n_{x'}$, the proportion of nodes in the network of prolificity x. Now consider the adjacencies: They number twice as many as the number of links. If we let $p_{x,w}$ be the proportion of adjacencies with author (node) prolificity x and collaboration (link) weight w then

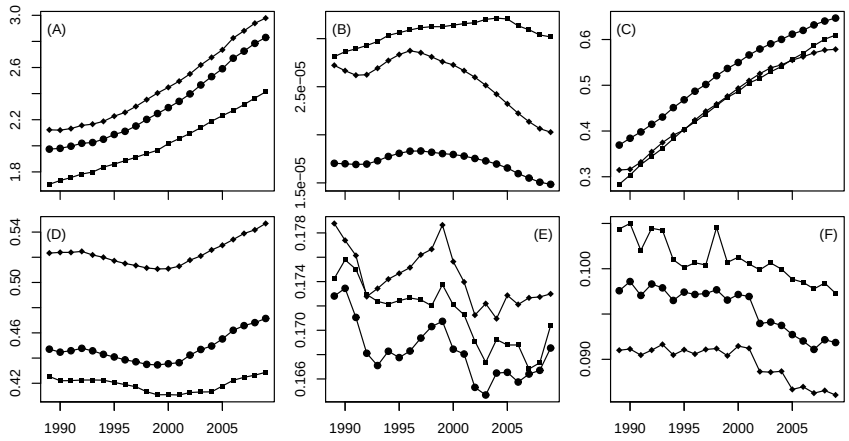

Figure S1: Across 5-year sliding windows: (A) Average number of authors on publications by one-time authors. One-time authors appeared on increasingly collaborative publications throughout our interval. (B) Network density. The applied network grew dramatically sparser after 1995 while the pure grew denser until 2006. (C) Proportion of nodes in the largest component. There is no difference between the pure and applied subnetworks to be discerned from the proportions of authors comprising their largest components, and the proportion comprising the largest component of the aggregate is consistently larger than both. (D–F) Proportions of authors with 1, 2, and 3 publications.

$q_r = \sum_{w \geq 1} p_{w+r,w}$ is the proportion of all adjacencies having remaining prolificity r.

Let us have ordered pairs (r, s) range over the remaining prolificities of linked nodes, so that each link is counted twice (as (r, s) and as (s, r)). Our statistic of interest is then

$$r_{\text{pub}} = \frac{E(rs) - E(r)E(s)}{\sqrt{Var(r)Var(s)}},$$

the correlation coefficient for the remaining prolificities of linked nodes.

Define $e_{r,s}$ to be the joint probability distribution of the remaining prolificities at the ends of a uniformly randomly chosen link. Since r and s are drawn from the same distribution, we may simplify the numerator as

$$\begin{aligned} E(rs) = E(r)E(s) &= E(rs) - E(r)^2 \\ &= \sum_r \sum_s rse_{r,s} - (\sum_r rq_r)^2 \end{aligned}$$

and the denominator as

$$\begin{aligned} \sqrt{Var(r)Var(s)} = Var(r) &= E(r^2) - E(r)^2 \\ &= \sum_r r^2 q_r - (\sum_r rq_r)^2. \end{aligned}$$

If we index the links by $i = 1, \ldots, m$ and (arbitrarily) label the remaining prolificities of their ends r_i and s_i then we may rewrite

$$\begin{aligned} \sum_r \sum_s rse_{r,s} &= \frac{1}{m} \sum_i r_i s_i, \\ \sum_r rq_r &= \frac{1}{2m} \sum_i (r_i + s_i), \\ \sum_r r^2 q_r &= \frac{1}{2m} \sum_i ({r_i}^2 + {s_i}^2). \end{aligned}$$

This provides the computational formula

$$r_{\text{pub}} = \frac{\frac{1}{m} \sum_i r_i s_i - (\frac{1}{m} \sum_i \frac{1}{2}(r_i + s_i))^2}{\frac{1}{m} \sum_i \frac{1}{2}({r_i}^2 + {s_i}^2) - (\frac{1}{m} \sum_i \frac{1}{2}(r_i + s_i))^2}.$$

If the remaining prolificities of linked nodes are independent then $e_{r,s} = q_r q_s$. If, instead, linked pairs are perfectly correlated in this respect then we get $e_{r,s} = q_r \delta_{r,s}$, where $\delta_{r,s}$ is the Kronecker delta (1 if $r = s$, 0 otherwise). The authors of a collaboration network are perfectly correlated by remaining prolificities r precisely when they are precisely correlated by prolificity x — that is, when the network consists of connected components of uniform prolificity.

8.3. Assortative mixing with low-count authors removed

To check that the trends and fluctuations we observed in r_{col} and in r_{pub} were not artifacts of the mixing behavior of authors with only one collaborator or publication, we ran the calculations on the aggregate with such authors removed from consideration. The overall trends were the same (Fig. S2).

8.4. Clustering coefficients

The time series for C (Fig. S3(A)) and r_{col} are similar. The dependence between these statistics reflects the reduced number of possible triangles among nodes of different degrees, which admits less clustering in disassortative graphs [SV05]. In the main paper we accounted for this interaction

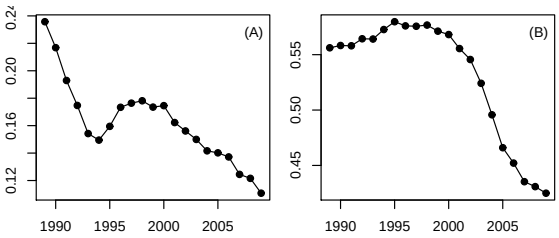

Figure S2: Across 5-year sliding windows on the aggregate network: (A) r_{col} calculated after removing authors with only one collaborator from consideration. (B) r_{pub} calculated after removing authors with only one publication from consideration.

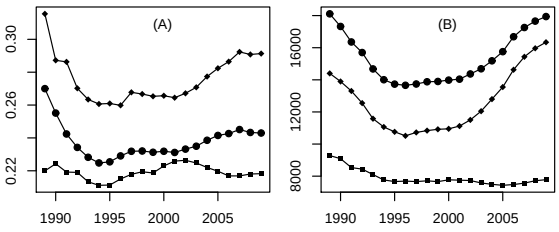

Figure S3: Across 5-year sliding windows: (A) Global clustering coefficient C: the proportion of connected triples of authors who are in fact pairwise linked. The time series closely resembles that of degree assortativity, particularly before 2001. (B) Clustering normalized by density (Fig. S1(D)), its expected value in a uniformly random graph, $C \cdot n(n-1)/2m$; the relative probability that two authors collaborated provided they had a common coauthor. The rises in density and in assortative mixing from 1995 to 2000 were largely to credit for the perceived rise in clustering during this period, while clustering after 2001 becomes more pronounced when corrected for these phenomena. Plots (B) and (C) use information from the entire graph.

using the correction $\tilde{C}$ introduced by Soffer and Vázquez [SV05].

Under uniformly random linking, a higher proportion of connected triples will form triangles in a denser graph, increasing clustering. To account for this, we normalized C by density (Fig. S3(D)) to get the *relative probability* that two authors have collaborated *provided that* they have a common coauthor [New01a] (Fig. S3(C)). Because density decreased substantially due to proportional growth in the largest component, this normalization increased monotonically across largest components; we took the normalization over entire graphs instead.

8.5. Change point fits

We fit change point models to time series data that were not clearly piecewise linear, but that exhibited one or two major changes in behavior amid smaller perturbations that the models interpret as normally-distributed error. Here we discuss change point models in more detail, and we display all aggregate residual plots, together with change point fits, used for the analysis.

The principle behind change point models is the same as that behind linear models. A traditional linear fit takes the form

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad \epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma),$$

while our change point model takes the form

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 (x_i - c)\delta_{x_i > c} + \epsilon_i, \quad \epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma).$$

We caution that some basic statistical assumptions for change point models are not met by this data: Particularly because adjacent sliding windows share 4 out of their 5 years but also because most authors publish in multiple years, measurements performed on these windows cannot be considered independent. Because each window contains a different (increasing) number of publications, they cannot be considered identically distributed. By performing change point analysis we do not intend to make predictions of future behavior but only to take advantage of an effective method for identifying shifts in behavior otherwise well-modeled linearly.

What follows is a simplification of the code we used to perform change point analysis in `R`. We required a guess `c` at the change point and calculated estimators for the coefficients by fitting a linear model `lm1` to the data below `c` (providing $\hat{\beta}_0$ and $\hat{\beta}_1$) and a linear model `lm2` with fixed intercept at $(\texttt{c}, \texttt{lm1(c)})$ to the data above `c` (providing $\hat{\beta}_2$).

```
# FUNCTION: Change point analysis on a
# collection of ordered pairs
changepoint.model <- function(
  x,    # points (independent), sorted
  y,    # values (dependent)
  c     # number in range(x)
) {
  len <- length(x)
  stopifnot(len == length(y))
  # Linear model to estimate b0 and b1
  m <- max(which(x < c))
  lm1 <- lm(y[1:m] ~ x[1:m])
  b0 <- lm1$coeff[1]
  b1 <- lm1$coeff[2]
  # Scaling model to estimate b2
  # y-value at x = c
  int <- lm1$coeff[1] + lm1$coeff[2] * c
  # x-values with origin (c,int)
  x2 <- x[(m + 1):len] - c
  # y-values with origin (c,int)
  y2 <- y[(m + 1):len] - int
  lm2 <- lm(y2 ~ x2 + 0)
  b2 <- lm2$coeff[1] - lm1$coeff[2]
```

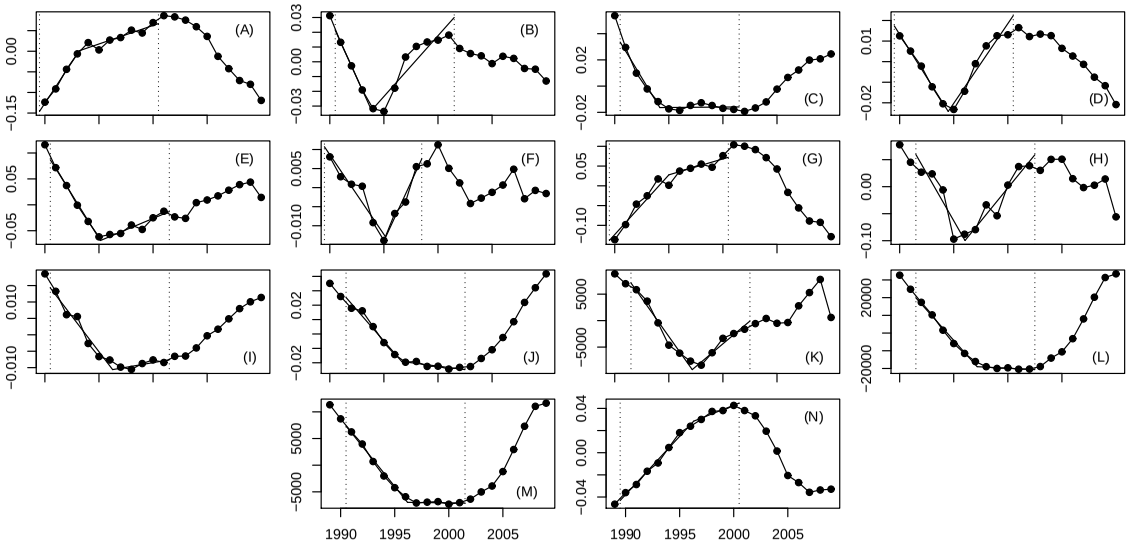

Figure S4: Residuals from best linear fits overlaid with a change point fit about the mid-90s event (5-year sliding windows).

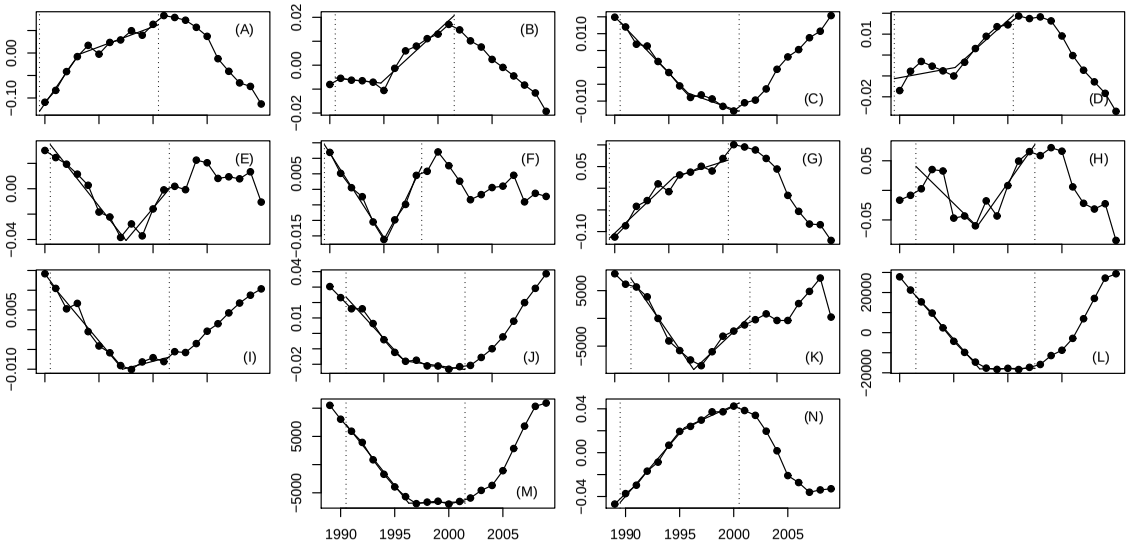

Figure S5: Residuals from best linear fits overlaid with a change point fit about the mid-90s event (5-year sliding windows, few-author network).

```
  # Change point model using estimators for
  # c (given), b0, b1, and b2
  return(summary(nls(
    as.formula(
      'y ~ B0 + B1 * x +
       B2 * (x - C) * (x >= C)'
    ),
    start = list(
      C = c, B0 = b0, B1 = b1, B2 = b2
    )
  )))
}
```

Fig. S4–S7 depict the change point fits we used to examine the two events in the main paper, with the exception of Fig. S4(Q); the fit, we judged, was too poor to warrant inclusion, and it demonstrates by comparison the superior fits obtained in other cases. In each plot the dotted vertical lines demarcate the intervals used to construct the model.

8.6. NSF funding for mathematics

The 1998 Odom Report [Odo98] recommended steep increases in funding for mathematics research, and from 2001 to 2004 annual NSF funding for its Division of Mathematical Sciences rose dramatically (Fig. S8). A change point fit to these numbers over 1995–2004 identifies a change year of 2000.36. Using 5-year averages instead, with each interval identified by its last year following the pattern used for other statistics, a change point fit over 1996–2006 identifies the change year 2001.33. Both values are toward the beginning of the collection of change years identified for network statistics, supporting a causal hypothesis, but significantly later than several specific statistics, suggesting that increased NSF funding may have contributed to, but was not the sole driver of, the second event.

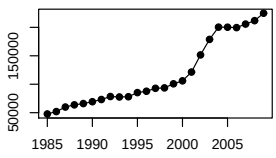

Figure S8: Funding by the National Science Foundation's Division of Mathematical Sciences, 1985–2007. A surge in funding beginning in 2001 was concurrent with the second event, specifically the surge in authorship, and leveled off after 2005.

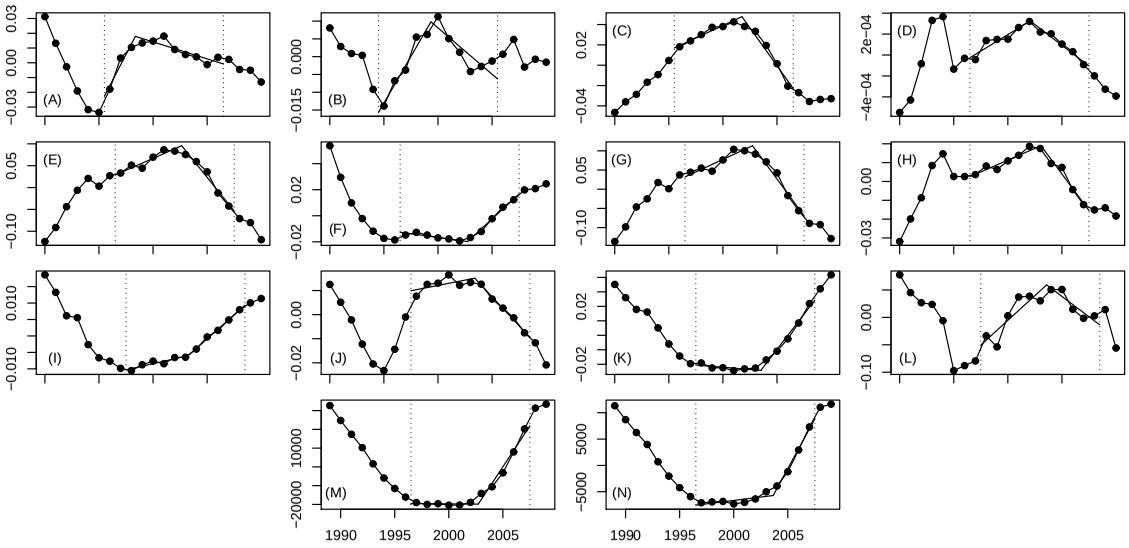

Figure S6: Residuals from best linear fits overlaid with a change point fit about the early-00s event (5-year sliding windows).

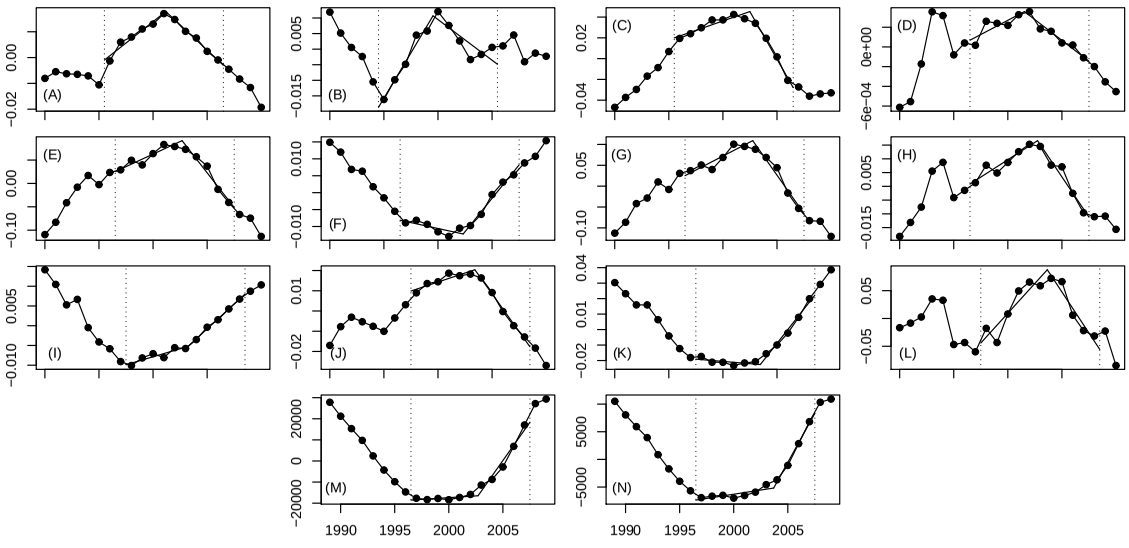

Figure S7: Residuals from best linear fits overlaid with a change point fit about the early-00s event (5-year sliding windows, few-author network).