Cooperation and coauthorship in scientific publishing

Lucas Wardil[*] and Christoph Hauert
Department of Mathematics, The University of British Columbia, 1984 Mathematics Road, Vancouver, British Columbia, Canada V6T 1Z2

(Received 12 August 2014; published 30 January 2015)

Research collaboration occurs more frequently today than in the past. As a consequence, cooperation and competition are crucial determinants of academic success. In multiauthored publications, not all authors contribute evenly. Hence, some authors end up with less time or resources to work on parallel projects, decreasing their number of publications. Although detailed information on the contribution of each author in multiauthored publications is generally not available, the order of authors often discloses information on differential contributions. Here we analyze the full data set of *Physical Review* journals to show that, along with the increasingly number of multiauthored publications, first authors incur costs and last authors are bestowed benefits in terms of number of publications. In other words, authors publishing more often as first authors have fewer publications in the short-term than authors publishing more often as last authors. Using a simplified network representation where direct links represent the costly action of first authors towards last authors, we analyze the evolution of cooperation in multiauthored publications.

DOI: 10.1103/PhysRevE.91.012825 PACS number(s): 89.75.−k, 89.65.−s, 02.50.Le, 01.30.−y

I. INTRODUCTION

Publishing is an important part of academic life, one that usually follows research project elaboration, securing funding, and labor-intensive times carrying out experiments or working on theory. Even though publishing is just making results publicly available, it is important for success in academic life [1–3]. Various indicators, like the number of publications and citation counts, have been proposed to measure success in academic life [4]. The h index, for instance, combines both citations and number of publications into a single number [5]. The large number of publications cataloged in digital libraries opened new possibilities for extensive analysis of how success builds. A key element to success nowadays is to develop a good network of collaborations. In the past, research was mainly done by solitary researchers. However, this scenario has drastically changed, mainly due to modernization of communication systems; increasing sophistication and cost of scientific instrumentation; as well as multidisciplinary projects, which require teams of experts in different fields [6]. The number of authors per publication has been increasing—with many individuals having only a few coauthors but a few having many coauthors [7,8]. Individual success is, therefore, becoming more dependent on successful collaborations, raising potential conflicts between cooperation and competition, for example, when different research groups working on the same topic are in a race for primary authorship [9].

Scientific collaboration refers to a broad range of activities, from simple opinion exchanges to side-by-side work in a laboratory. In any case, collaborators recognized as authors have their names acknowledged on the article's byline (the line in article's head displaying author names). The recognition of authors' contributions in physics and life science is usually indicated by the hierarchical ordering of authors. Statistical analysis of the author order on the byline reveals that young researchers are usually first authors, and more senior researchers are usually last authors, while not so much can be said regarding middle authors [10]. However, this is not the rule in other fields, including mathematics, economics, and high-energy physics, where author names are usually alphabetically ordered [11]. Nonetheless, if author names are not alphabetically ordered, the byline order undoubtably provides some information about the underlying division of labor.

In multiauthored publications, authors contribute with different types and amounts of resources, increasing the number of publications of all coauthors equally. Ideally, one would like to have as many collaborators as possible contributing with the more costly resources, so parallel projects can be executed [12]. However, this raises the traditional dilemma of cooperation: Who is willing to pay the costs [13]? Cooperative behavior has been widely studied within the framework of evolutionary game theory, both theoretically and experimentally [14–19]. In behavioral sciences, cooperation is usually defined as an action where the actor incurs costs to generate benefits for the group. The actor may, or may not, enjoy the benefits. In dyadic interactions, the former is referred to as a snow-drift game and the latter as a prisoner's dilemma game [20]. In biology, costs and benefits of a behavior are measured in terms of the effect on the number of offspring: Costly behavior decreases the number of offsprings and beneficial behavior increases it [21]. In behavioral economics, costs and benefits are usually represented by monetary incentives [22]. Similarly, in multiauthored publications, costs and benefits associated with research strategies can be measured in terms of the number of publications. Hence, this poses the question of how specific strategies affect author productivity.

Research strategies are complex traits that depend on individual capabilities and preferences as well as on environmental contingencies, like availability of funding or technologies, administrative assignments, and so on. In publications a myriad of individual decisions and environmental conditions are reflected in the order of authors in the publication's byline. Here we show that the position on the byline is strongly correlated with the number of individual publications, as evinced by the analysis of the entire data set of *Physical Review*

[*] Corresponding author: wardil@math.ubc.ca

journals. While multiauthoring has obviously increased the overall number of publications published in science, such an increase comes at a short-term cost for first authors, who on average end up with fewer publications than last authors. Using a simplified network model, the cooperative act where the first author transfers a benefit to the last author in terms of publication output can be represented as a directed link from the first to the last author. In this way, the analysis of network topology provides information about cooperation pattern evolution in scientific publishing.

II. DATA-SET ANALYSIS

The data set consists of publications in *Physical Review* journals, hereafter referred to as *PR* journals, from 1893 to 2009 (see Appendix). To provide a more continuous approach to data, we first divide the data set into 5-year blocks consisting of all publications between years t and $t + 4$, with $1893 \leqslant t \leqslant 2005$. In each 5-year block an author publishes N publications, where N_s, N_f, N_m, and N_l are the number of publications where the author is single, first, middle (in any position), or last author, respectively. Note that $N_s + N_f + N_m + N_l = N$. Therefore the strategy profile of an author in each 5-year block can be represented by (x_s,x_f,x_m,x_l), where $x_i = N_i/N$ with $i \in \{s,f,m,l\}$.

The overall fraction of multiauthored publications in *PR* journals has drastically increased since the first publication of *Physical Review* in 1913, from 30% to roughly 96% of multiauthored publications in the 5-year block starting in 2005, Fig. 1. To analyze the relation between the number of publications and author position on byline, we group authors with the same number N of publications in a given 5-year block. Then we calculate the average $\bar{x}_i$, for $i \in \{s,f,m,l\}$, for each group. The graphs of $(\bar{x}_i,N)$ in the 5-year block starting in 2000 are shown in Fig. 2. The graph of $(\bar{x}_s,N)$, Fig. 2(a), shows that the fraction of single-authored publications is indeed small, slightly lower for large N. On the other hand the graphs of $(\bar{x}_f,N)$ and $(\bar{x}_l,N)$ suggest that authors with higher productivity are less likely to work as first authors and more likely to work as last authors. To quantify these claims, we measured the correlation between N and $\bar{x}_i$, i.e., the correlation between number of publications of an author and average fraction of publications in each position for the given author.

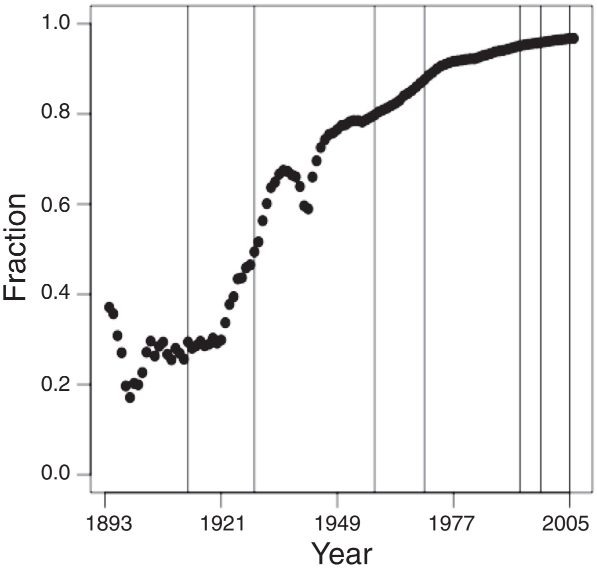

FIG. 1. Fraction of multiauthored publications (black) published in *Physical Review* journals from 1983 to 2005. Each year corresponds to a 5-year block. Vertical lines represent the date of the first publication of the following *Physical Review* journals: *Phys. Rev.* (1913), *Rev. Mod. Phys.* (1929), *Phys. Rev. Lett.* (1958), *Phys. Rev. A, B, C,* and *D* (1970), *Phys. Rev. E* (1993), *Phys. Rev. Spec. Top-AB* (1998), *Phys. Rev. Spec. Top-AC* (2005).

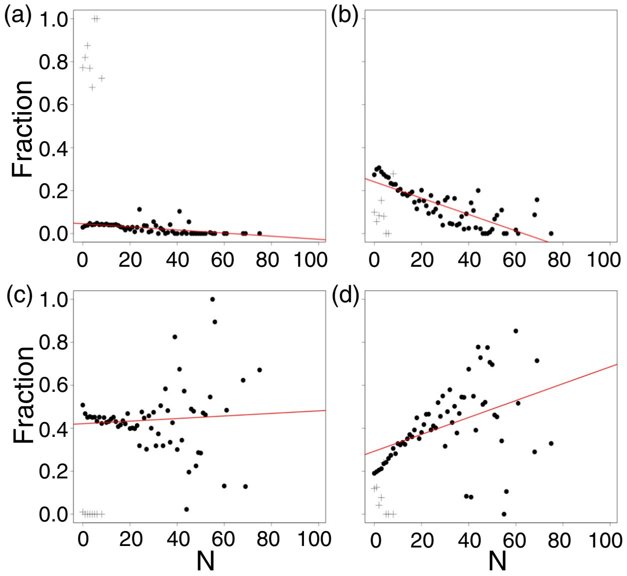

FIG. 2. (Color online) Fraction of publications published in each byline position averaged over authors with the same number of publications in the 5-year block starting in 1900 (crosses) and in 2000 (filled circles). For authors publishing N publications, the figure shows (a) the average fraction of publications as single author, $\bar{x}_s$; (b) the average fraction as first author, $\bar{x}_f$; (c) the average fraction as middle author, $\bar{x}_m$; and (d) the average fraction as last author, $\bar{x}_l$. To illustrate the trend, linear regression of the points in the 5-year block starting in 2000 is represented by the solid red line.

To measure correlation we use the Kendall's τ coefficient, a nonparametric statistics that, for a set $\{(y_1,z_1),(y_2,z_2),\ldots\}$, essentially counts the number of times that $y_k - y_s$ and $z_k - z_s$ have the same sign for all k, s. The correlation analysis generates a single number, the Kendall's τ coefficient ($-1 \leqslant \tau \leqslant 1$). Positive τ values indicate positive correlation, and negative values indicate negative correlation. For example, the correlation coefficients for $(\bar{x}_s,N)$, $(\bar{x}_f,N)$, $(\bar{x}_m,N)$, and $(\bar{x}_l,N)$ in the 5-year block starting in 2000 are given by -0.57^*, -0.68^*, -0.03, and 0.45^*, respectively (significant values—p value < 0.01—are superscripted with $*$). In 2000 the number of publications is positively correlated with $\bar{x}_l$ and negatively correlated with both $\bar{x}_s$ and $\bar{x}_f$.

The correlation coefficients of $(\bar{x}_i,N)$ for $i \in \{s,f,m,l\}$ across the entire *PR* journal lifespan are shown in Fig. 3. The correlation analysis indicates that, roughly, since the launch of

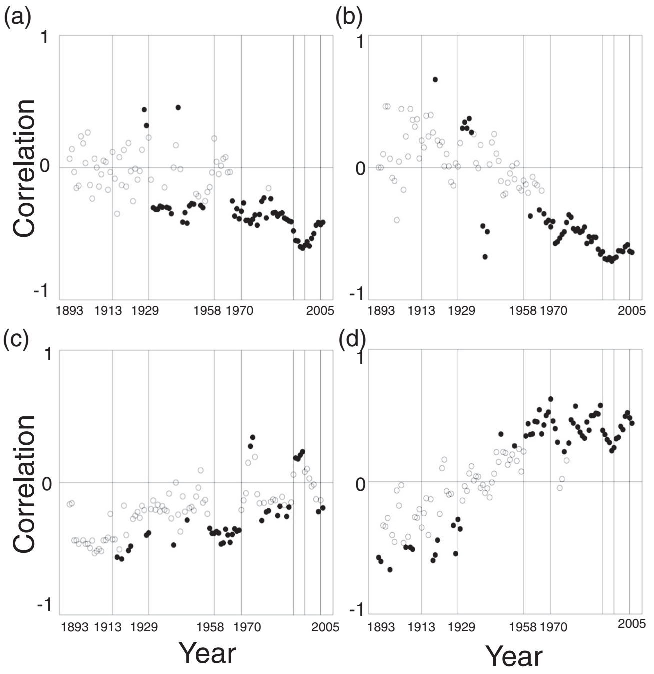

FIG. 3. Evolution of correlation between the total number of publications of an author and the average fraction of publications in which authors are (a) single, (b) first, (c) middle, and (d) last authors. Correlation is given by Kendall's τ coefficient. Data displayed as filled dots are statistically significant with p value < 0.01.

Rev. Mod. Phys. in 1929 there is a positive correlation between productivity and fraction of last-authored publications. It also indicates that since the launch of *Phys. Rev. Lett.*, in 1958, there is a negative correlation between productivity and first-authored publications (correlations are significant with p value < 0.01). Correlation between productivity and the fraction as middle authors does not support confident conclusions, with correlation values for most of recent years being not statistically significant.

These results show that, alongside the proliferation of multiauthored publications, working as first author has a negative effect on the publication rate, and working as last authors has a positive effect on it. Therefore, there must be costs incurred by first authors and benefits bestowed on last authors such that the overall effect is to lower the production rate of first authors compared to last authors. Hence, our analysis suggests that coauthorship relations are built upon a backbone of cooperative acts, giving rise to a dynamic network of cooperation.

III. COOPERATION NETWORK

Coauthorship interactions can be represented by a dynamical directed network [23], where authors are represented as nodes and the cooperative action of the first author is represented by a directed link pointing from the first author to the last author. In this way a directed link represents an action where donors, i.e., first authors, incur costs to provide benefits to the recipients, i.e., last authors. In this model the network structure encodes the individual cooperative behavior. Note that because the relative costs and benefits of middle author position cannot be clearly defined, only the relation between first and last authors are represented as directed links. The behavioral type of author i at the 5-year block starting at year t can be characterized by

$$L_i(t) = \frac{k_i(t) - l_i(t)}{k_i(t) + l_i(t)},$$

where $k_i(t)$ is the number of outgoing links of node i and $l_i(t)$ is the number of incoming links of node i. Note that the number of outgoing and incoming links represents the number of publications where the node is first and last author, respectively. Hence the number of outgoing links indicates the amount of costs incurred to the author and the number of incoming links indicates the amount of benefits bestowed on the author. The behavioral type L_i can be classified into three categories. If $-1 \leqslant L_i < -1/3$, the author contributes more as last author, managing to have others to work as first authors, and should be classified as *laird*. If $-1/3 \leqslant L_i \leqslant 1/3$, the author contributes as first and as last author to the same extent and should be classified as *trader*. If $1/3 < L_i \leqslant 1$, the author contributes more as first author and should be classified as *worker*. Defining the activity of author i at time t as

$$A_i(t) = k_i(t) + l_i(t),$$

the behavioral type of an individual can be represented in the phenotype space $L \times A$. Note that activity as defined here represents only the number of publications where the author is either first or last author.

The phenotype space and network snapshots for a few sample years are shown in Fig. 4. Networks at the beginning of the 20th century were very sparse because of few collaborations. With the increase of scientific collaboration worldwide, networks started to get denser with a large spread in the phenotype space. The frequency of *workers* is slightly increasing, Fig. 5(a), likely due to the increasing number of students in science. The average cooperation level of each behavioral type remained almost constant over time, as shown in Fig. 5(b). Authors adopting trader behavior have larger activity A_i, as can be seen in Fig. 5(c). Interestingly only after 1940 *laird* authors started to publish more than *worker* authors. This suggests that before the Second World War working as first author was actually not so costly, as indicated by some positive correlation values in Fig. 3(b).

Despite last authors ending up with more publications, this option is not available to everyone. The analysis of *PR* journal reveals that the option of working as last author is related to careers progress, accordingly to previous findings indicating a junior and a senior pattern of publication [10]. To quantify this claim we looked at the productive lifetime of an author, which can be approximated as the period between the first and the last publication [24,25]. Lifetime can then be split into three parts of equal length. In each third we calculate the average cooperation level L of each author. The histogram of behavioral types—*worker*, *trader*, or *laird* types—in each third is shown in Fig. 6. This result suggests that in early stages of their career authors act more as *workers*, while in late stages authors act more as *lairds*. We considered here only well-established authors, that is, only authors whose productive life span exceeds 15 years. As an example, the evolution of L for

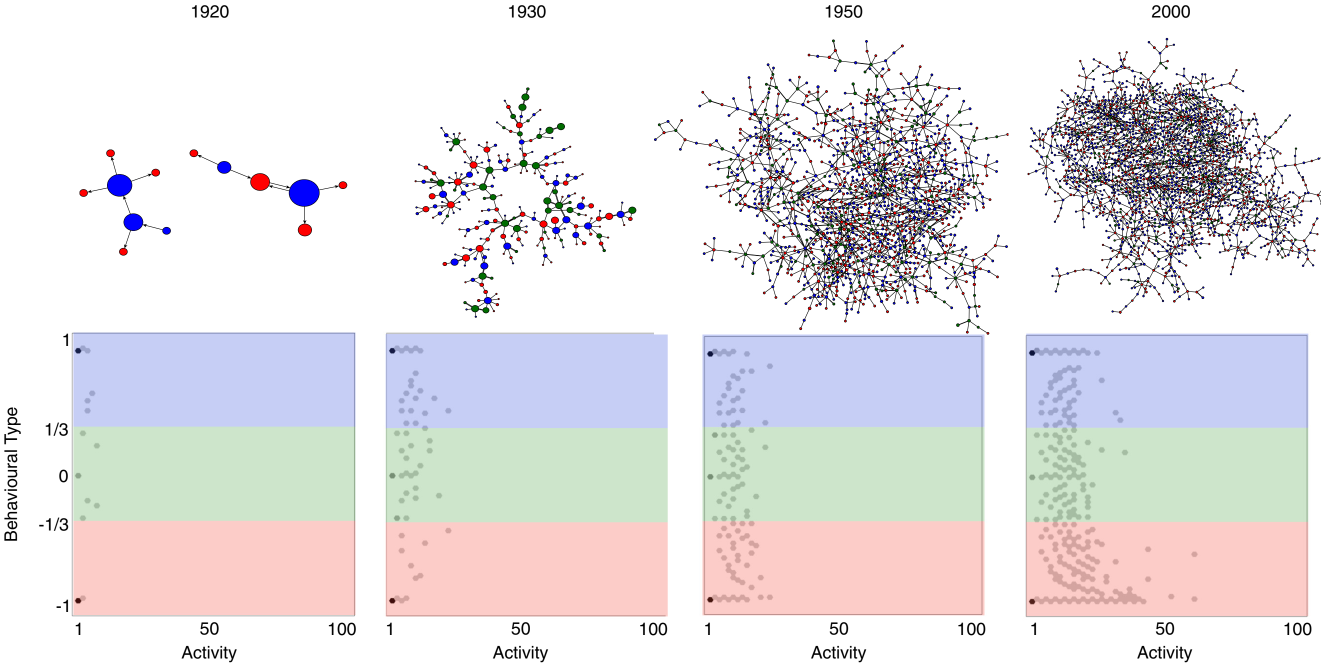

FIG. 4. (Color online) Emergence of social networks: sample snapshots of the largest connect clusters (top row) and the corresponding phenotype distribution (bottom row, histogram) for 1920, 1930, 1950, and 2000. The size of the nodes (different scales are used in each year) in the networks indicates individual's activity, N_i, and the color its behavioral type: *workers* (blue, $1/3 < L_i \leqslant 1$), *traders* (green, $-1/3 \leqslant L_i \leqslant 1/3$), and *lairds* (red, $-1 \leqslant L_i < -1/3$). The network for 2000 is too big and dense to be plotted. Therefore only the largest connected component is shown with 90% of links randomly removed.

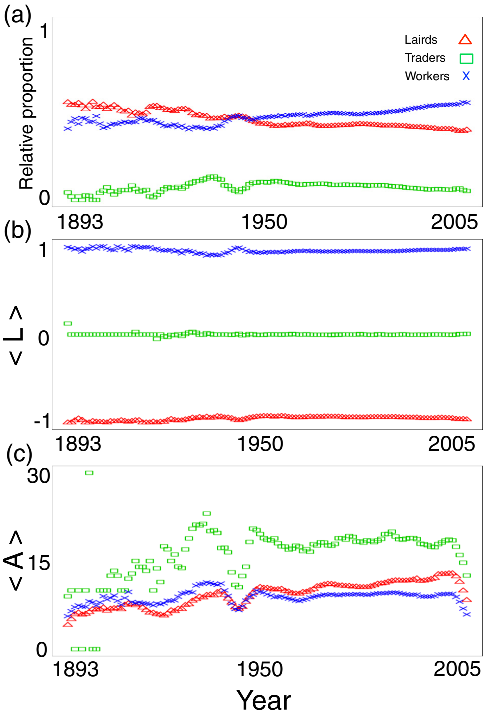

FIG. 5. (Color online) Average behavioral type evolution of *workers* (blue crosses), *traders* (green squares), and *lairds* (red triangles). (a) Evolution of the relative proportion of each type in each 5-year block. Evolution of the average values of (b) cooperation level L_i and (c) activity A_i considering each type as a sample. Note that there is a sharp decrease in activity around 1940 due to the Second World War.

the top four authors of 2000, Fig. 7, shows that authors start as *workers* and become *lairds* as they advance in their career.

IV. DISCUSSION AND CONCLUSION

In our study, we argue that being the first author in a multiauthored publication represents an act of cooperation because it bestows benefits of additional publications on the last author at costs to first author. Cooperation in multiauthored papers is much like the snow-drift game metaphor, where one driver—the first author—pays the cost of shovelling snow to clear the road, while the other—the last author—stays inside the car [26]. Although both enjoy the benefits of going home, one pays higher costs. The detailed characterization of costs and benefits of research collaboration is, however, a hard task. For example, while some authors may contribute with the execution of time-consuming tasks, other authors may contribute with provision of funds or research infrastructure.

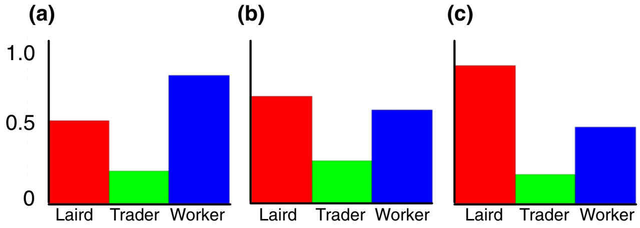

FIG. 6. (Color online) Probability density distribution of behavioral types in (a) the first third of the career, in (b) the second third, and in (c) the last third for each behavioral type.

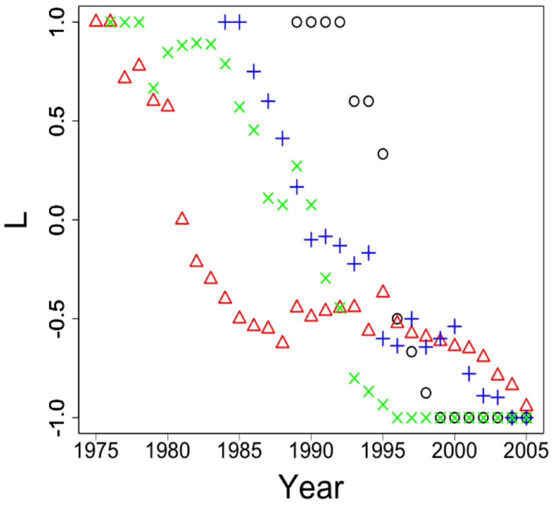

FIG. 7. (Color online) Evolution of the cooperation level L_i for the four most productive authors in 2000 through their publication lifetime in *PR* journals. Each author is represented by a different symbol.

In both cases, authors may incur costs in terms of their number of publications. Regardless of other factors, our analysis shows that byline position is an indirect indicator of costs and benefits. Given that authors at the beginning of their careers are not at the point of being the main providers of research funds and institutional support, our analysis suggests that the costs are related to the time to carry out the research, which prevents first authors to take on as many parallels projects as last authors.

An alternative way to formulate the costs and benefits involved in a multiauthored publications is in terms of opportunity costs. Opportunity cost is the value of the best alternative forgone [27]. In the data set, the fractions of first- and single-authored publications, x_f and x_s, respectively, are negatively correlated to the total number of publications N, whereas x_l is positively correlated to N, as shown in Fig. 3. Individuals increase their productivity, on average, by increasing n_l rather than n_s or n_f. Hence, if a researcher works as a first author instead of last author, the forgone benefit of extra publications represent his or her opportunity costs. Last authors naturally benefit from the collaboration by having more time to work on parallel projects. Hence, multiauthored papers represent an intrinsic act of cooperation in terms of publication output in that first authors pay opportunity costs to provide benefits to last authors. Note that opportunity costs are individual dependent. For example, if a well-connected researcher decides to focus on a large solitary project, the number of publications that she is likely to forfeit is much higher than if she was a not well-connected researcher.

Although being a *laird* can be highly beneficial in terms of the number of publications, sometimes, like at the beginning of a career, the only strategic movement one may have is to add directed links, i.e., to be a *worker* and produce publications as first author. However, high-quality publications and fruitful collaborations with highly ranked authors may aggregate value to publications and help to promote the author to a status where more incoming links can be attracted.

Activity A_i should not be confounded with the total number of publications, as it does not include middle authored publications. Although Fig. 3(c) suggests that most of the time the correlation between fraction as middle author and productivity is negative, this result is not statistically significant. Hence the phenotype space $L \times A$ refers only to the backbone of cooperative actions in coauthorship networks where costs and benefits can be clearly assigned. Also note that, although single authors clearly incur costs, as indicated by the negative correlation in Fig. 3(a), it does not directly involve any coauthorship interaction. Thus only social interactions are accounted for in the phenotype space $L \times A$.

Evaluation of author productivity within *PR* journals database is just an approximation, as authors do not publish only in *PR* journals. The actual productivity of an author can be better approximated if we restrict the data set to include only authors that publish a large number publications per year in *PR* journals. With this stronger restriction, the claim that being first authors is costly and being last author is beneficial is even more supported by our data analysis (not shown).

To conclude, the analysis of *PR* journal publication data provides a unique opportunity to interpret coauthorship collaboration in terms of cooperative behavior taking place in a natural environment, as opposed to behavioral experiments in artificial laboratory setups. Here we argue that the relation between first and last author is built upon a cooperative component, where the first author pays a short-term cost in terms of number of publications and the last author receives a short-term benefit. Such a relationship is easily mapped to a dynamical model of directed networks, where cooperative actions are represented by the network topology. Because *lairds* and *workers* are the majority types since the launching of the first *PR* journal in 1893, unveiling the actual individual contributions in multiauthored publications is crucial to a fairer assessment of author productivity.

ACKNOWLEDGMENTS

This research was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) as well as the Foundational Questions in Evolutionary Biology Fund (FQEB, John Templeton Foundation), Grant No. RFP-12-10.

APPENDIX: DATA SET

The data set consists of 462 090 publications in *Physical Review* journals from 1893 to 2009. Each entry corresponds to a single publication, containing information on publication date, an ordered list of authors, and other pieces of information. Although each publication has a unique identifier, there are some ambiguities concerning author names [28]: (i) the same author uses different byline names or (ii) different authors use the same byline name. To decrease the first type of ambiguity, we associate each byline name to a key composed by the last name and the initials of the byline name. For example, the byline names "L. Wardil" and "Lucas Wardil" have the same key "LW:Wardil." If two, and only two, different byline names are associated to the same key, we assume that the same author is using two different byline names. In this case the key is used as author identifier. In the original database 17% of byline names pertain to this first class. However, if more than two byline names are associated to the same key, we assume that more than one author are sharing the same key. In this case the full byline name is used as author identifier. In the original database 15% of byline names pertain to this second class. To fix this second type of ambiguity, more information is required, like affiliation and coauthorship patterns, and still many false positives remain [4]. Because we look only at the number of publication within a short moving time frame of 5 years, the second type of ambiguity is minimal and will not be treated in our analysis.

To infer author contribution from author order, publications alphabetically ordered should be excluded. Alphabetical order may arise by chance in publications with few authors, but in publications with many authors this is not the case. A preliminary analysis of the *PR* data set reveals that for more than five authors the fraction of publications with authors alphabetically ordered is significantly larger than random. Hence, as a rule of thumb, we exclude publications with more than five alphabetically ordered authors. The filtered data set contains 348 487 publications and 219 142 authors, in contrast to the original number of 460 889 publications and 235 533 authors.

[1] G. J. Feist, Psychol. Sci. **4**, 366 (1993).
[2] M. F. Fox and P. E. Stephan, Soc. Stud. Sci. **31**, 109 (2001).
[3] N. Carayol and M. Matt, Res. Policy **33**, 1081 (2004).
[4] A. M. Petersen, F. Wang, and H. E. Stanley, Phys. Rev. E **81**, 036114 (2010).
[5] J. E. Hirsch, Proc. Natl. Acad. Sci. USA **102**, 16569 (2005).
[6] M. Bordons and I. Gómez, in *The Web of Knowledge: a Festschrift in Honour of Eugene Garfield*, edited by B. Cronin and H. B. Atkins (Information Today, Inc., American Society for Information Science, Medford, 2000), p. 197.
[7] T. Martin, B. Ball, B. Karrer, and M. E. J. Newman, Phys. Rev. E **88**, 012814 (2013).
[8] M. E. J. Newman, Proc. Natl. Acad. Sci. USA **101**, 5200 (2004).
[9] P. Atkinson, C. Batchelor, and E. Parsons, Sci. Technol. Hum. Val. **23**, 259 (1998).
[10] R. Costas and M. Bordons, Scientometrics **88**, 145 (2011).
[11] T. V. Frandsen and J. Nicolaisen, J. Informetr. **4**, 608 (2010).
[12] M. O. Jackson and A. Wolinsky, J. Econ. Theor. **71**, 44 (1996).
[13] R. Axelrod, *The Evolution of Cooperation* (Basic Books, New York, 1984).
[14] J. A. Fletcher and M. Doebeli, Proc. R. Soc. B **276**, 13 (2009).
[15] G. Szabó and G. Fáth, Phys. Rep. **446**, 97 (2007).
[16] L. Wardil and J. K. L. da Silva, Europhys. Lett. **86**, 38001 (2009).
[17] D. Semmann, Proc. Natl. Acad. Sci. USA **109**, 12846 (2012).
[18] J. Grujic *et al.*, Sci. Rep. **4**, 4615 (2014).
[19] C. Camerer, *Behavioral Game Theory: Experiments in Strategic Interaction* (Princeton University Press, Princeton, NJ, 2003).
[20] J. A. Fletcher and M. Doebeli, J. Evol. Biol. **19**, 1389 (2006).
[21] T. Day and S. P. Otto, Fitness, in *eLS* (Wiley & Sons Ltd., Chichester, 2001).
[22] A. M. Colman, *Game Theory and Its Applications in the Social and Biological Sciences* (Butterworth-Heinemann, London, 1995).
[23] L. Wardil and C. Hauert, Sci. Rep. **4**, 5725 (2014).
[24] J. C. Huber, Scientometrics **45**, 33 (1999).
[25] J. Duch, X. H. T. Zeng, M Sales-Pardo, F. Radicchi, S. Otis, T. K. Woodruff, and L. A. N. Amaral, PLoS ONE **7**, e51332 (2012).
[26] M. Nowak, *Evolutionary Dynamics: Exploring the Equations of Life* (Havard University Press, Cambridge, MA, 2006).
[27] W. A. McEachern, *Microeconomics: A Contemporary Introduction* (South-Western Cengage Learning, Mason, OH, 2014).
[28] F. Radicchi, S. Fortunato, B. Markines, and A. Vespignani, Phys. Rev. E **80**, 056103 (2009).