Model-based evaluation of scientific impact indicators Supporting Information

Matúš Medo[1, *] and Giulio Cimini[2, 3, †]

[1] *Physics Department, University of Fribourg, CH-1700 Fribourg, Switzerland*
[2] *IMT School for Advanced Studies, 55100 Lucca, Italy*
[3] *Istituto dei Sistemi Complessi (ISC)-CNR, 00185 Rome, Italy*

(Dated: September 7, 2016)

PAPER CITATION COUNT VERSUS PAPER APPEARANCE TIME

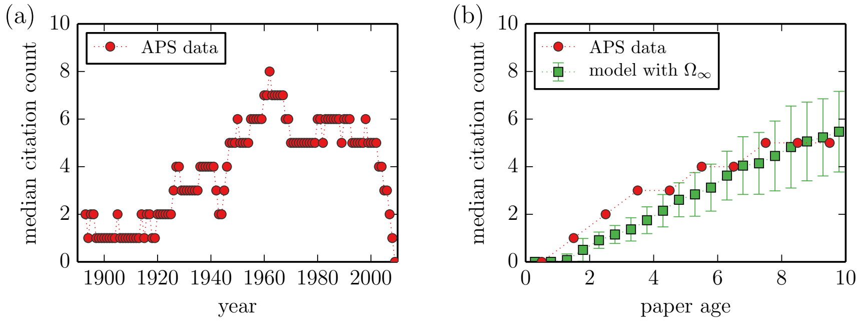

Figure S1. (a) The dependence of the median paper citation count on the publication year in the standard American Physical Society (APS) data that cover all papers published by the APS journals and citations among them (the covered period is 1893–2009). The citation count decrease for papers published after 2000 can be attributed to these papers not having time to reach their long-term impact. The citation count decrease for papers published before 1950 can be attributed to different citations practices (fewer references per paper). Unlike the model without Ω_∞, the data feature no median peak for the earliest published papers.
(b) A comparison of the dependence of the median paper citation count in the model data with Ω_∞ (the basic model setting; the error bars show three-fold of the standard error of the mean computed from 100 model realizations) and the APS data. Given the fact that the model was not calibrated with respect to this measurement, the two curves agree remarkably well.

[*] matus.medo@unifr.ch
[†] giulio.cimini@imtlucca.it

VARIATIONS OF THE BASIC SETTING

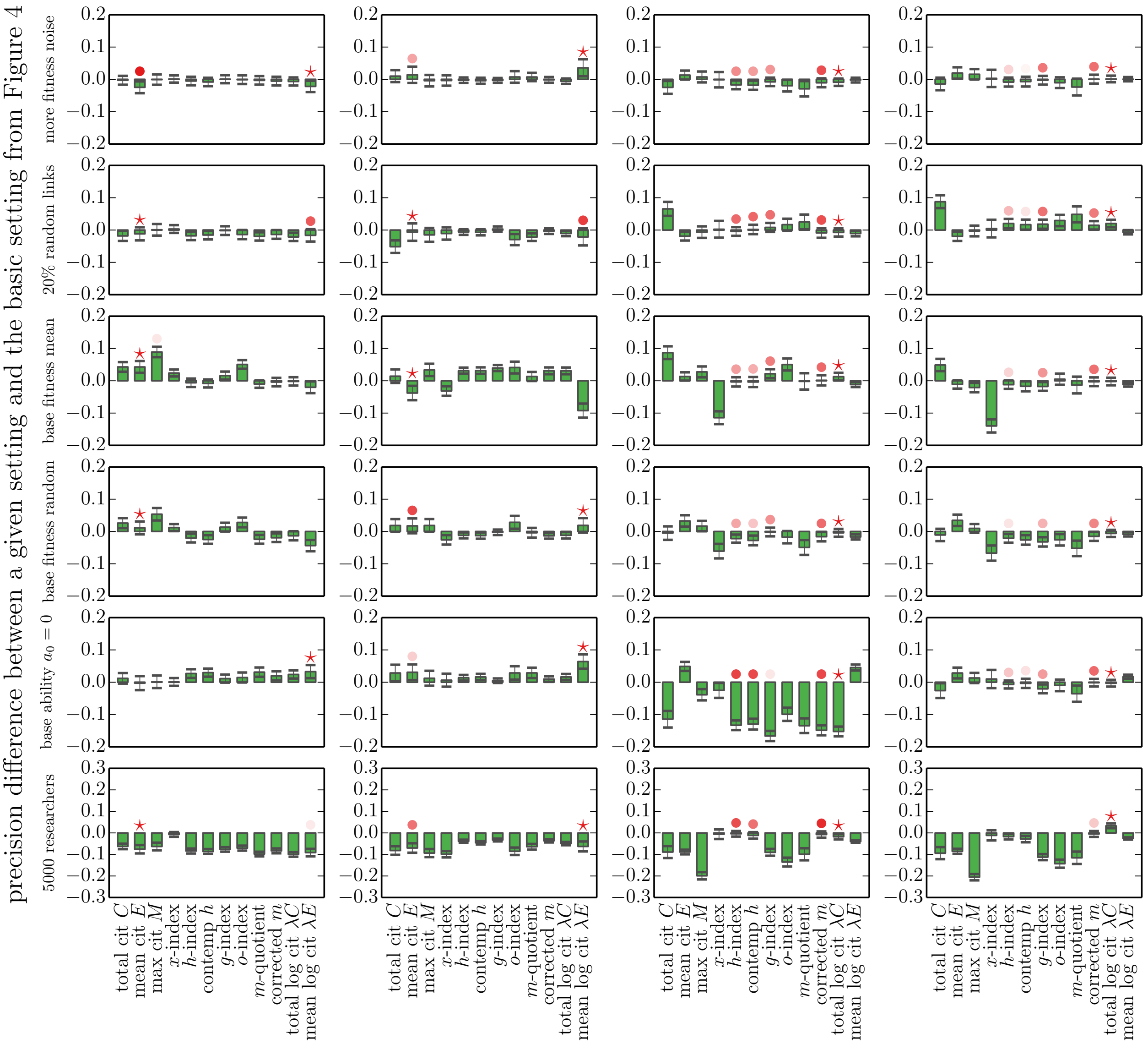

Figure S2. Precission difference with respect to the basic setting for individual metrics when various variations of this setting are considered (from top to bottom):
(1) the fitness noise amplitude η^* is increased from 0.2 to 0.5 ($\Omega_\infty = 1.48 \cdot 10^4$),
(2) 20% of references made by a new paper target a random existing paper ($\Omega_\infty = 1.28 \cdot 10^4$),
(3) the base fitness of a paper is obtained as the average of the authors' ability values ($\Omega_\infty = 9.20 \cdot 10^3$),
(4) the base fitness of a paper is a random sample from the authors' ability values ($\Omega_\infty = 1.10 \cdot 10^4$),
(5) the baseline ability of authors a_0 is set to zero ($\Omega_\infty = 1.06 \cdot 10^4$),
(6) the number of researchers is increased from 1000 to 5000 ($\Omega_\infty = 6.90 \cdot 10^4$).
The different ground truth assumptions are, from left to right: author ability, average fitness of authored papers, author ability times activity, and total fitness of authored papers. The error bars show the three-fold of the standard error of the mean. For every setting and ground truth, the best-performing metric (in absolute terms) is marked with a star ($\star$), and all metrics that reach at least 90% of its precision are marked with bullets ($\bullet$) whose color indicates the performance difference ($\bullet$, $\bullet$, and $\bullet$ correspond to 98%, 95%, and 92% of the best performance, respectively).

AUTHORS AT DIFFERENT CAREER STAGES

Here we study the case where new authors are gradually introduced to the system, and check which indicators allow to fairly evaluate young researchers with respect to their senior colleagues. We thus consider a situation where 25% of researchers are active for the whole time; for the remaining 75%, author i begins their activity in a randomly chosen step τ_i between $t = 1$ and $t = T - 36$. Here 36 is subtracted to only include researchers who spent sufficient time in the system (by doing so, we are essentially considering only researchers who finished their PhD studies). In this setting, the productivity of author i, initially drawn from $\mathcal{G}(k)$, is linearly rescaled by her appearance time in the system by a factor $(1 - \tau_i/T)$: young authors are thus in disadvantage with respect to seniors by having on average less publications at the end of the simulation, and also by having less time to accrue citations.

Figure S3 presents the results obtained with this setting. The first observation is that most metrics actually perform similarly than in the basic setting where the group of active researchers does not change over time. The only metric that considers authors' career length, m-quotient, is surprisingly performing worse in the new setting with researchers gradually entering the system. The reason lies again in the rescaling problem mentioned in the main text: in the new setting, there are more young users who only author their papers in the last years and yet outperform venerable researchers upon the rescaling, thus lowering the resulting precision more than in the basic setting presented in Figure 4 of the main text. Specifically, there are on average 28 ± 8 authors with only one publication in the top 100 positions of the ranking by the m-quotient and the average activity span of top 100 researchers is 60 ± 40 months (out of 120 in total). By contrast, for the h-index ranking there are no authors with only one publication in top 100 and the average activity span of top researchers is 96 ± 21 months. The bias towards very young researchers is removed by the corrected m, which brings to no researchers with only one paper in top 100 and to an average activity span of 86 ± 28 months: the resulting precision is similar to that achieved with the h-index and the contemporary h-index. Overall, the logarithm-based indices λE and λC are again the best performers against intensive and extensive ground truths, respectively.

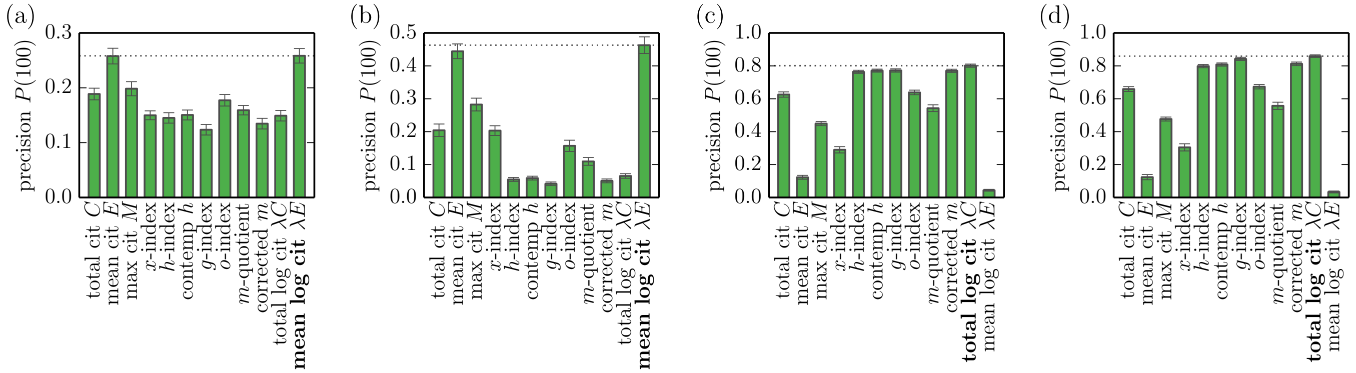

Figure S3. For the setting when the number of active researchers grows in time (see the description in the text above), comparison of the mean precision values achieved by impact indicators with respect to different ground truth assumptions: (a) author ability, (b) average fitness of authored papers, (c) author ability times activity, (d) total fitness of authored papers. The horizontal dotted line marks the performance of the best metric (which is typed with bold letters); the error bars show three-fold of the standard error of the mean.

INTERMEDIATE GROUND TRUTH ASSUMPTIONS

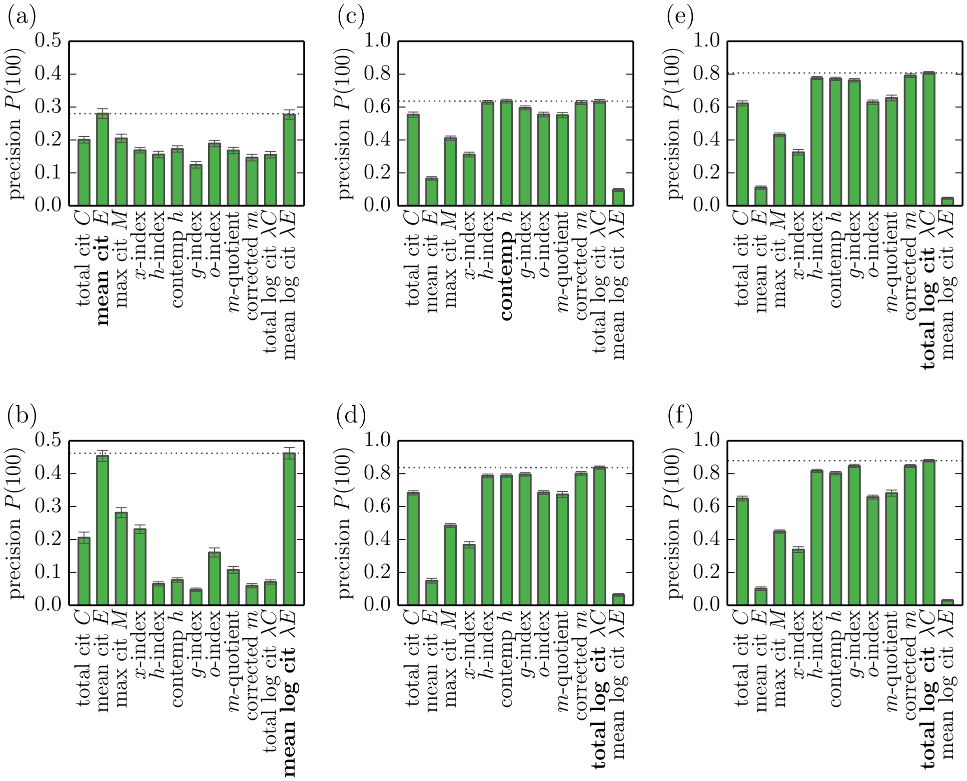

Figure S4. Precision values achieved by individual metrics against intensive (a,b), midway (c,d) and extensive (e,f) ground truth assumptions: (a) author ability, (b) author ability times square root of activity, (c) author ability times activity, (d) average fitness of authored papers, (e) average fitness of authored papers times square root of activity, (f) average fitness of authored papers times activity—*i.e.*, total fitness of authored papers. The first row of panels (a,c,e) refers to a benchmark obtained from the researchers' potential, whereas, the second row (b,d,f) to a benchmark related to realized publication outputs. We assume here the basic simulation setting that was used to obtain Figure 3 in the main text, with $\Omega_\infty = 1.33 \cdot 10^4$.

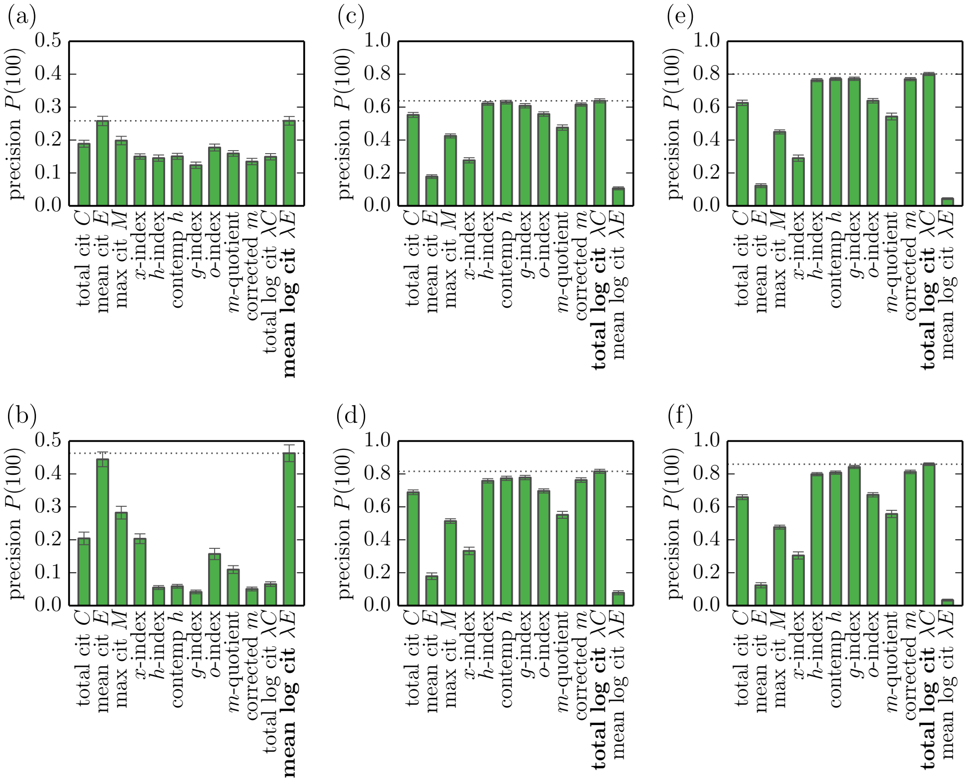

Figure S5. Same as Figure S4, but for the simulation setting with the number of active researchers increasing with time as in Figure S3. Here $\Omega_\infty = 1.10 \cdot 10^4$.