Supporting Information

Defining and Identifying Sleeping Beauties in Science

Qing Ke, Emilio Ferrara, Filippo Radicchi, and Alessandro Flammini

Center for Complex Networks and Systems Research,

School of Informatics and Computing,

Indiana University, Bloomington, Indiana 47408, USA

(Dated: March 30, 2015)

S1. DATASETS

In this work, we use two large datasets, namely the American Physical Society (APS) and the Web of Science (WoS). APS contains $463,348$ papers published from 1893 to 2009 in APS journals and is publicly available upon request at http://journals.aps.org/datasets; WoS is comprised of $35,174,034$ papers published between 1900 and 2011 in journals covering most research fields, and is available upon purchase from Thomson Reuters. Most papers in the APS dataset are also in the WoS. The APS dataset, though, contains fewer citations: only those originating from papers within the APS journals are therein recorded. Our analysis is based on papers that received at least one citation. A total number of $384,649$ and $22,379,244$ such papers were found in the APS and WoS dataset, respectively. Fig. S1 shows the yearly number of papers with at least one citation received before the end of the observation period. The fact that recent papers have had less time to accumulate citations is reflected in the sharp decrease that is noticeable as time approaches the end of the observation period.

S2. EXAMPLES OF TOP SLEEPING BEAUTIES

Figs. S2 and S3 show the citation history of the top 24 papers in the APS dataset. Table S1 presents the comparison between our results and Redner's results [8].

Fig. S4 displays the citation history of the top 15 Sleeping Beauties in the WoS dataset showed in Table I of the main text. Tables S2, S3, and S4 present the basic information of the top Sleeping Beauties in Statistics, Mathematics, and Social Sciences and Humanities, respectively. See Figs. S5–S8 for corresponding citation histories.

S3. CHARACTERIZING DECREASING PATTERNS

This section presents a statistical characterization of how yearly citations of papers decrease after the peak. In summary, for most of the papers the yearly citation rate decreases quickly (possibly exponentially) after its peak. Our analysis focused only papers with positive beauty coefficient B, for a total of $189,673$ (out of $384,649$; 49.3%) and $14,689,643$ (out of $22,379,244$; 65.6%) papers in the APS and WoS dataset, respectively. We further classify every of these papers into two categories depending on whether or not their yearly citation counts c_t decreased to half of its maximum during the observation period $[t_m+1, T]$ (Figs. S9*A*-*B*).

We identify 18,131 (9.56%) papers in the APS whose c_t have not decreased below $c_{t_m}/2$, and 2,094,671 (14.26%) in the WoS dataset. Figs. S9*C*–*D* display the histograms of $T - t_m$. We observe that a large fraction are recently awakening papers, with about 60% of them getting their maximum yearly citations c_{t_m} in the last year of the observation periods ($T - t_m = 0$).

For the remaining papers whose yearly citations have decreased below $c_{t_m}/2$, we define the paper "half-life" t_h as the number of years required by c_t to decrease from c_{t_m} to $c_{t_m}/2$. Figs. S9*E*–*H* show the distributions of t_h across all these papers in the APS (Fig. S9*E*), papers whose B values ranked in the top 1% (Fig. S9*F*), from 1% to 10% (Fig. S9*G*), and the rest (Fig. S9*H*). We see that yearly citations of SBs decrease rapidly after the peak regardless of their B values. These results are confirmed also in the WoS dataset, as shown in Figs. S9*I*–*L*.

S4. NULL MODELS

To verify that the beauty coefficients cannot be explained by the underlying citation networks or other well-known mechanisms, we compare the citation history of each paper as well as the beauty coefficient distribution with those obtained from some null models. Here we employ two null models on the APS dataset, namely citation network randomization (NR) and the preferential attachment mechanism (PA).

The NR procedure starts from the original citation network and carries out a series of link swapping. The end-point nodes (the papers being cited) of a randomly selected pair of links (citations) are swapped if: (i) the two links do not share source or target node; (ii) there are no multiple links after swapping; and, (iii) the publication year of the cited article is not greater than that of the citing article after swapping. Performing $Q \cdot E$ switches, where E is the number of links in the citation network and Q is set to 50, yields a transformation of the original citation network into a random directed graph. This procedure preserves for each paper its number of references (out degree) and total number of citations (in degree), but destroys the dynamics of yearly citations.

PA considers as initial network the empirical APS citation network from 1893 to 1897 when the first citation occurred; it contains 182 nodes and 1 link. In each following year t until 2009, n_t papers are added at the same time, and each paper p brings r_p references. n_t is set to the number of APS papers actually published in year t and each r_p corresponds to the number of references of one of the papers in such set. As we progressively add papers to the citation network, the references they contain are addressed to previously published papers chosen with probability proportional to one plus the number of citations those papers already have.

S5. COARSE TOPICS OF SLEEPING BEAUTIES IN THE APS

Examining the citation relationships between papers with high B values gives us some coarse topics of Sleeping Beauties. In Fig. S10 we present the citation network of the 100 papers with the highest B values in the APS dataset. Despite many isolated nodes, we observe some (weakly) connected components. Diving into each component, we find that each one corresponds to one coarse topic. In Fig. S11, for instance, we show the topic of each of the 4 largest components and the citation histories of its constituent papers. Except for Fig. S11(b), we observe that papers belonging to the same group exhibit remarkably similar citation histories. They are awoken in the same year and exhibit similar up- and down-going citation patterns. Fig. S11(a) shows the double exchange mechanism works. This theory was introduced in 1950s and became popular in the 1990s. The second group shown in Fig. S11(b) is about Quantum Mechanics. The central paper (blue line and blue node), which is cited by every other paper in the group, is the famous EPR paradox paper by Einstein, Podolsky, and Rosen. The third group shown in Fig. S11(c) is particularly interesting, as it exhibits complex fluctuations in the citation histories. Finally, the group shown in Fig. S11(d) is about graphite and graphene. The central paper (blue line and blue node) in Fig. S11(d) is a pioneering work on the band structure of graphite, foundation of the discovery of graphene, the subject of the 2010 Nobel Prize in Physics.

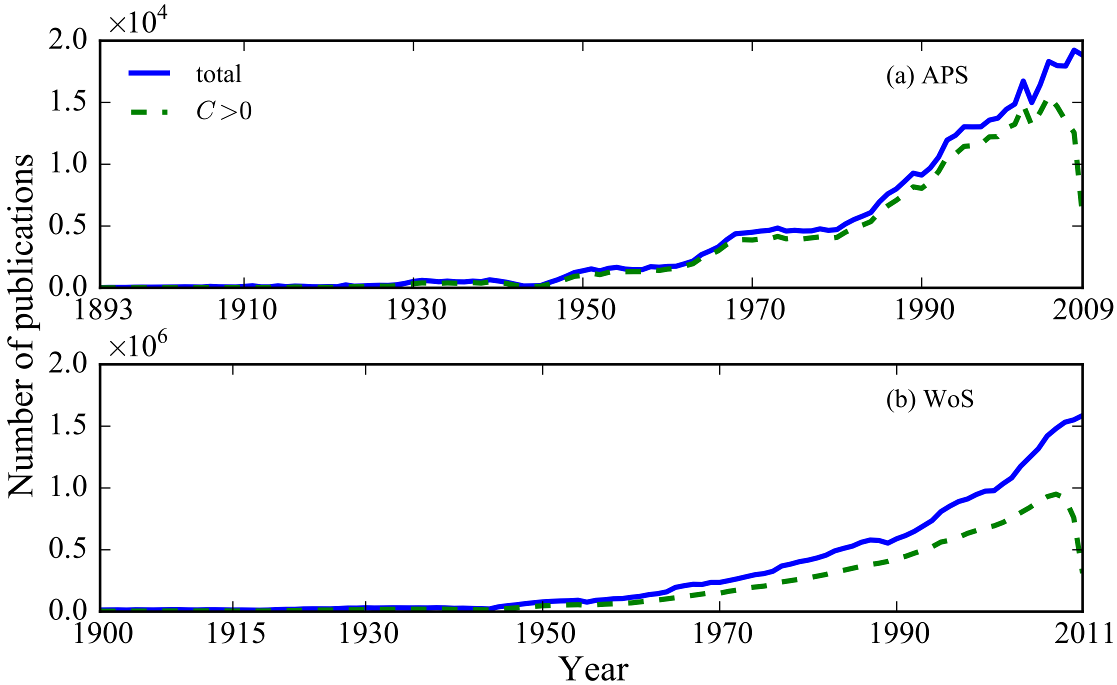

FIG. S1. (Blue solid) Total number of papers per year; (Green dashed) Yearly number of papers that received citations.

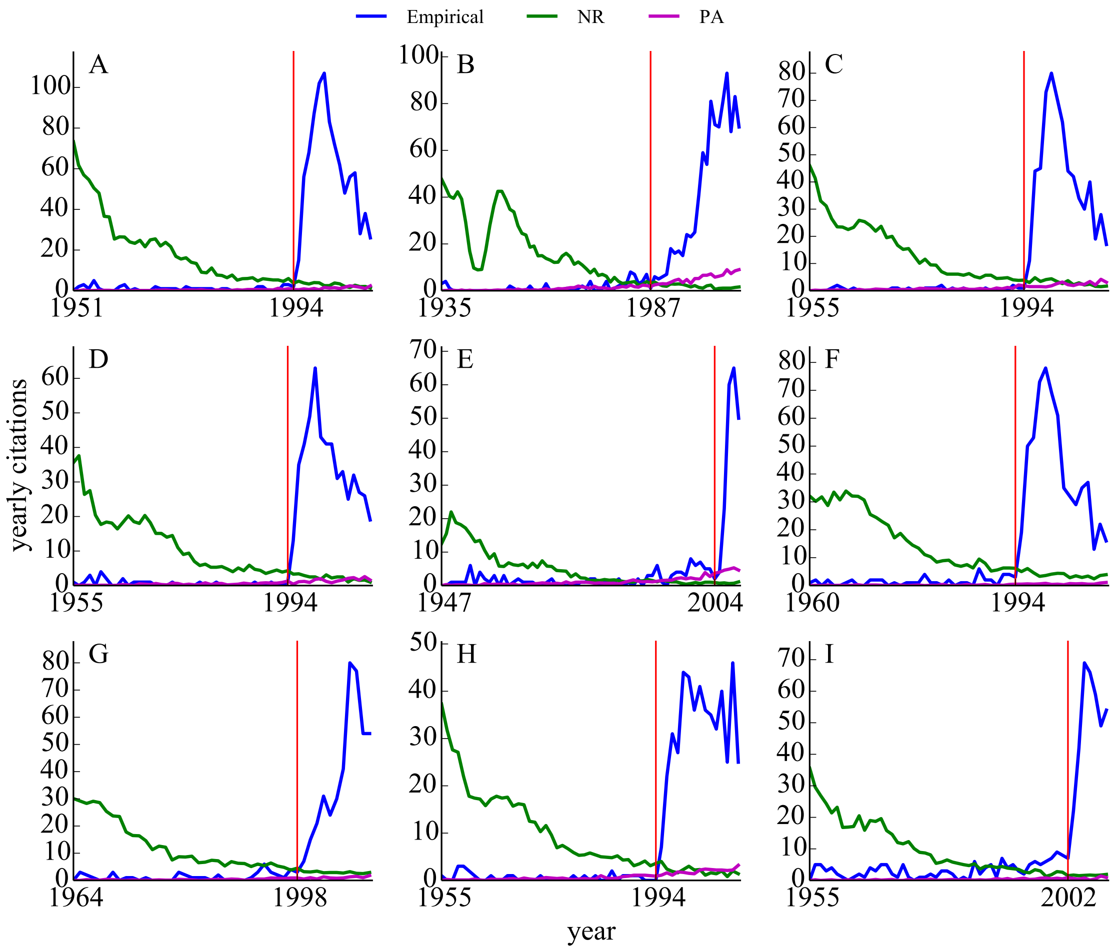

FIG. S2. Top Sleeping Beauties in physics. Blue curves show yearly citations received by papers: (A) Phys. Rev. 82, 403 (1951), $B = 1,722$ [12]; (B) Phys. Rev. 47, 777 (1935), $B = 1,419$ [4]; (C) Phys. Rev. 100, 675 (1955), $B = 1,348$ [1]; (D) Phys. Rev. 100, 545 (1955), $B = 1,107$ [10]; (E) Phys. Rev. 71, 622 (1947), $B = 1,086$ [9]; (F) Phys. Rev. 118, 141 (1960), $B = 841$ [2]; (G) Phys. Rev. 135, A550 (1964), $B = 825$ [5]; (H) Phys. Rev. 100, 564 (1955), $B = 670$ [7]; (I) Phys. Rev. 100, 580 (1955), $B = 624$ [3]. Yearly citations obtained from citation network randomization (NR) and preferential attachment (PA) model are plotted as green and purple lines, respectively. Both the NR and PA results are averaged across 10 realizations. The awakening years, identified using Eq. 3, are indicated by the vertical red lines. The sharp decrease of the curve for the NR result in panel B is probably due to the decrease of number of publications during the period of World War II (Fig. S1a). Panels A, C, D, F, and H refer to papers about the double exchange mechanism. Panel B refers to the EPR paradox paper by Einstein, Podolsky, and Rosen. Panel E considers the pioneering study on the band structure of graphite.

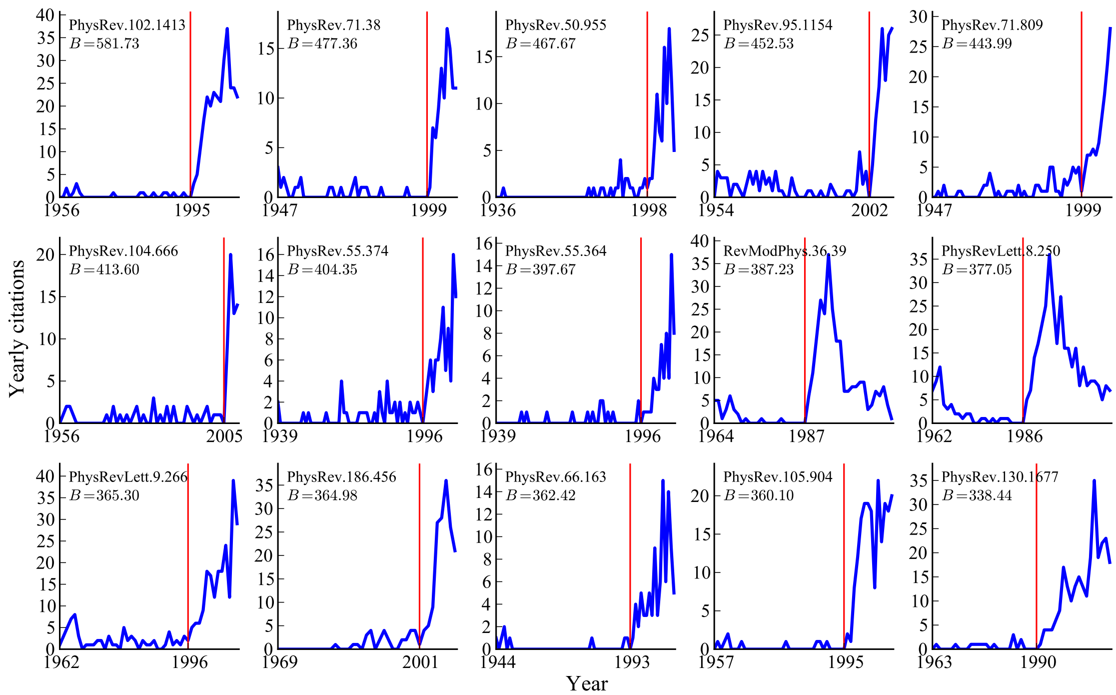

FIG. S3. (Blue) Citation histories, (Red) awakening years, and B values of the 15 papers ranked from 10^{th} to 24^{th} based on the B values in the APS dataset. The ending year is 2009.

Publication	Rank	B	Awakening
PR 40, 749 (1932)	45	250.79	1980
PR 46, 1002 (1934)	54	237.40	1975
PR 47, 777 (1935)	2	1419.15	1987
PR 56, 340 (1939)	96	174.59	1987
PR 82, 403 (1951)	1	1722.25	1994
PR 82, 664 (1951)	192	122.56	2007
PR 100, 545 (1955)	4	1106.82	1994
PR 100, 564 (1955)	8	670.42	1994
PR 100, 675 (1955)	3	1348.26	1994
PR 109, 1492 (1958)	147	138.63	2004
PR 115, 485 (1959)	218	115.07	2001
PR 118, 141 (1960)	6	841.47	1994

TABLE S1. Comparison between our results and Redner's results [8]. The first column lists the 12 *revived classics* in physics detected by Redner's analysis and arranged in chronological order. From the second column, we report our results: the rank position according to their beauty coefficient B, the value of B, and the awakening year.

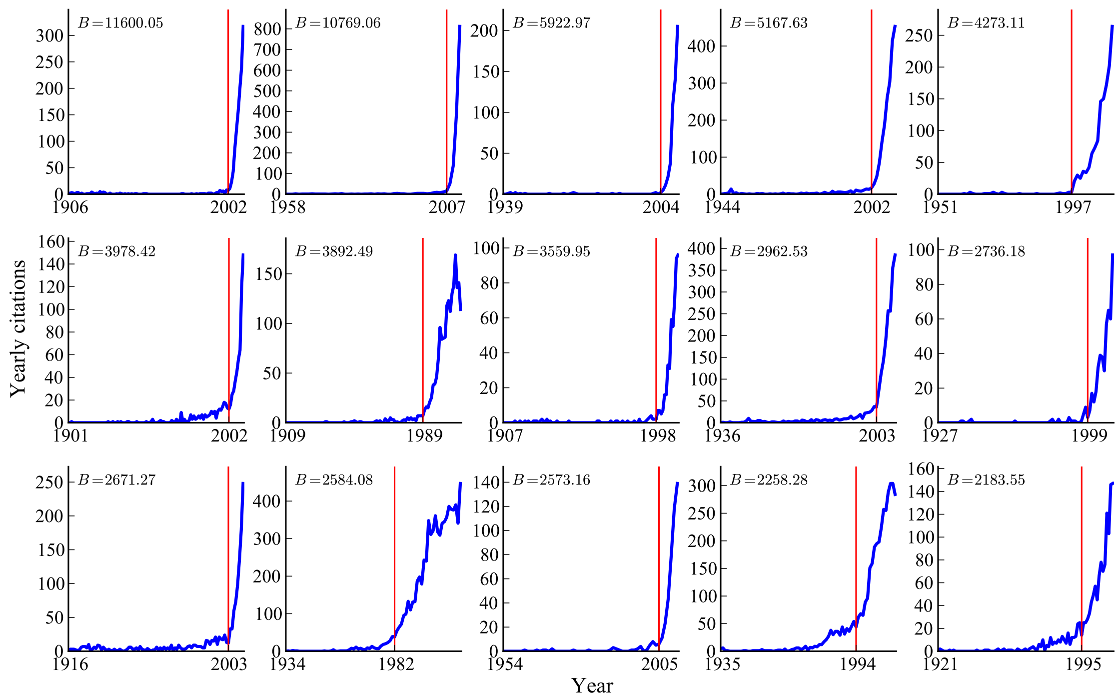

FIG. S4. (Blue) Citation histories, (Red) awakening years, and B values of the top 15 papers, based on the B values in the WoS dataset. The ending year is 2011.

B	Author	Title	Pub., awake	Journal
3978	Pearson, K	On lines and planes of closest fit to systems of points in space	1901, 2002	Philos. Mag.
2736	Wilson, EB	Probable inference, the law of succession, and statistical inference	1927, 1999	J. Am. Statist. Assoc.
1909	Mann, HB	Nonparametric tests against trend	1945, 2003	Econometrica
1893	Kaplan, EL; Meier, P	Nonparametric estimation from incomplete observations	1958, 1980	J. Am. Statist. Assoc.
1760	Fisher, RA	On the interpretation of χ^2 from contingency tables, and the calculation of P	1922, 2006	J. R. Stat. Soc.
1247	Hastings, WK	Monte-carlo sampling methods using markov chains and their applications	1970, 1995	Biometrika
1193	Metropolis, N	The monte carlo method	1949, 2004	J. Am. Statist. Assoc.
1124	Moran, PAP	Notes on continuous stochastic phenomena	1950, 1999	Biometrika
1050	Lorenz, MO	Methods of measuring the concentration of wealth	1905, 2005	J. Am. Statist. Assoc.
985	Kendall, MG	A new measure of rank correlation	1938, 2004	Biometrika

TABLE S2. Basic information about the top 10 papers in Statistics. See Fig. S5 for their citation histories.

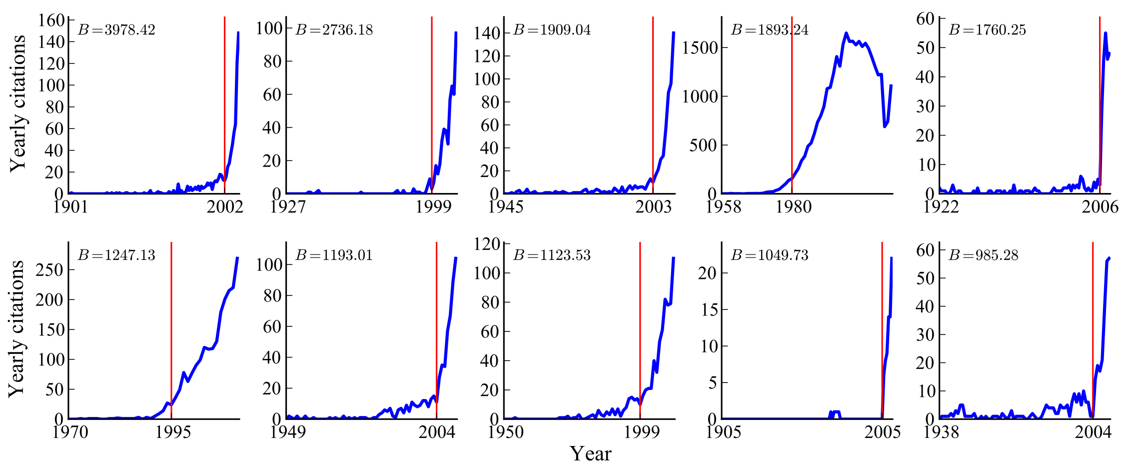

FIG. S5. (Blue) Citation histories, (Red) awakening years, and B values of top 10 papers in Statistics based on the B values in the WoS dataset. The ending year is 2011.

B	Author	Title	Pub., awake	Journal
1215	Wiener, N	The homogeneous chaos	1938, 2001	Amer. J. Math.
1060	Leray, J	On the movement of a viscous fluid to fill the space	1934, 1995	Acta Math.
851	Pringsheim, A	On the theory of the double infinite numerical orders	1900, 2005	Math. Ann.
765	Jensen, JLWV	On the convex functions and inequalities between mean values	1906, 2006	Acta Math.
706	Mann, WR	Mean value methods in iteration	1953, 2004	Proc. Am. Math. Soc.
670	Halpern, B	Fixed points of nonexpanding maps	1967, 2004	Bull. Amer. Math. Soc.
669	Haar, A	On the theory of orthogonal function systems (first announcement)	1910, 1988	Math. Ann.
609	Weyl, H	The asymptotic dispersal law of eigen values of linear partial equations differential (with an application for the theory of cavity radiation)	1912, 2002	Math. Ann.
578	Painleve, P	About second order and higher order differential equations whose general integral is uniform	1902, 1990	Acta Math.
558	Schmidt, E	On the theory of linear and non-linear integral equations chapter i development of random functions in specific systems	1907, 1992	Math. Ann.

TABLE S3. Basic information about the top 10 papers in Mathematics. See Fig. S6 for their citation histories.

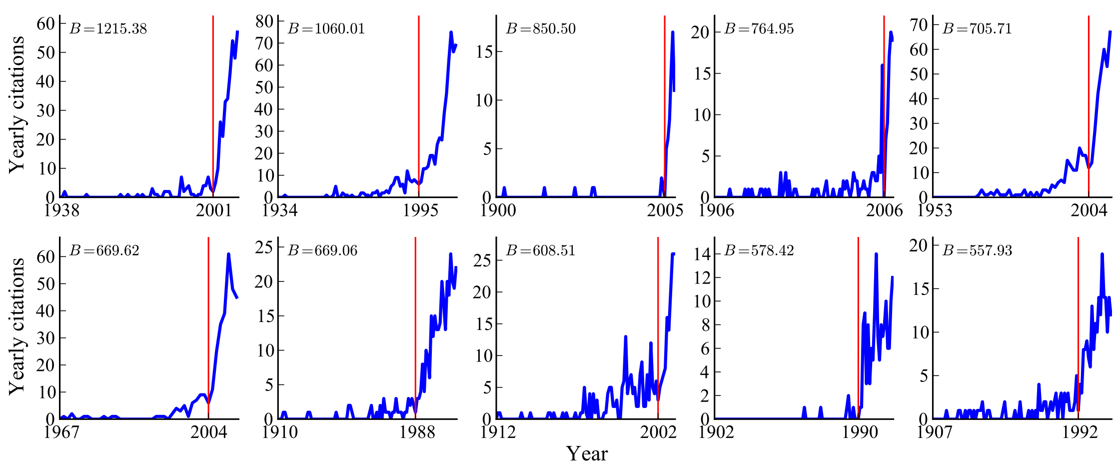

FIG. S6. (Blue) Citation histories, (Red) awakening years, and B values of top 10 papers in Mathematics based on the B values in the WoS dataset. The ending year is 2011.

B	Author	Title	Pub., awake	Journal
1901	Stroop, JR	Studies of interference in serial verbal reactions	1935, 1987	J. Exp. Psychol.
1255	Yerkes, RM; Dodson, JD	The relation of strength of stimulus to rapidity of habit-formation	1908, 1981	J. Comp. Neurol.
584	Zachary, WW	Information flow model for conflict and fission in small groups	1977, 2005	J. Anthropol. Res.
563	Tobler, WR	Computer movie simulating urban growth in Detroit region	1970, 2003	Econ. Geogr.
545	Garfield, E	Citation indexes for science - new dimension in documentation through association of ideas	1955, 2000	Science
545	Heider, F; Simmel, M	An experimental study of apparent behavior	1944, 1998	Am. J. Psychol.
521	Watson, JB	Psychology as the behaviorist views it	1913, 1968	Psychol. Rev.
488	Cohen, J	A coefficient of agreement for nominal scales	1960, 2009	Educ. Psychol. Meas.
485	Maslow, AH	A theory of human motivation	1943, 1998	Psychol. Rev.
479	Glaser, BG	The constant comparative method of qualitative analysis	1965, 2004	Social Problems
467	Todd TW	Age changes in the pubic bone	1921, 2003	Am. J. Phys. Anthropol.
460	Forrester, JW	Industrial dynamics - a major breakthrough for decision makers	1958, 1993	HBR
453	Rosenblatt, F	Perceptron - a probabilistic model for information storage and organization in the brain	1958, 2001	Psychol. Rev.
446	Hotelling, H	Analysis of a complex of statistical variables into principal components	1933, 1994	J. Educ. Psychol.
428	Thorndike, EL; Woodworth, RS	The influence of improvement in one mental function upon the of efficiency other functions (I)	1901, 1992	Psychol. Rev.
424	Holzinger, KJ; Swineford, F	The bi-factor method	1937, 2003	Psychometrika
405	Thistlethwaite, DL; Campbell, DT	Regression-discontinuity analysis - an alternative to the ex-post-facto experiment	1960, 2005	J. Educ. Psychol.
399	Horn, JL	A rationale and test for the number of factors in factor-analysis	1965, 2000	Psychometrika
375	Fisher, I	The debt-deflation theory of great depressions	1933, 2004	Econometrica
369	Spitzer, HF	Studies in retention	1939, 2004	J. Educ. Psychol.
368	Linn, BS; Linn, MW; Gurel, L	Cumulative illness rating scale	1968, 1999	J Am Geriatr Soc.
358	Hull, CL	The goal gradient hypothesis and maze learning	1932, 2001	Psychol. Rev.
356	Elftman, H; Manter, J	Chimpanzee and human feet in bipedal walking	1935, 2001	Am. J. Phys. Anthropol.
349	Fornell, C; Larcker, DF	Evaluating structural equation models with unobservable variables and measurement error	1981, 2004	J. Marketing Res.
343	Armstrong, JS; Overton, TS	Estimating nonresponse bias in mail surveys	1977, 1998	J. Marketing Res.
342	Wechsler, H	Toward neutral principles of constitutional-law	1959, 1986	Harv. Law Rev.
324	Cohen, J	Eta-squared and partial eta-squared in fixed factor anova designs	1973, 2005	Educ. Psychol. Meas.
324	Dunlap, K	Reactions to rhythmic stimuli, with attempt to synchronize	1910, 1995	Psychol. Rev.
320	Ellsberg, D	Risk, ambiguity, and the savage axioms	1961, 2002	Q. J. Econ.
320	Lewin, K	Defining the 'field at a given time'	1943, 2006	Psychol. Rev.

TABLE S4. Basic information about the Sleeping Beauties in Social Sciences and Humanities among the top $1,000$ in the WoS dataset. See Fig. S7 and S8 for their citation histories.

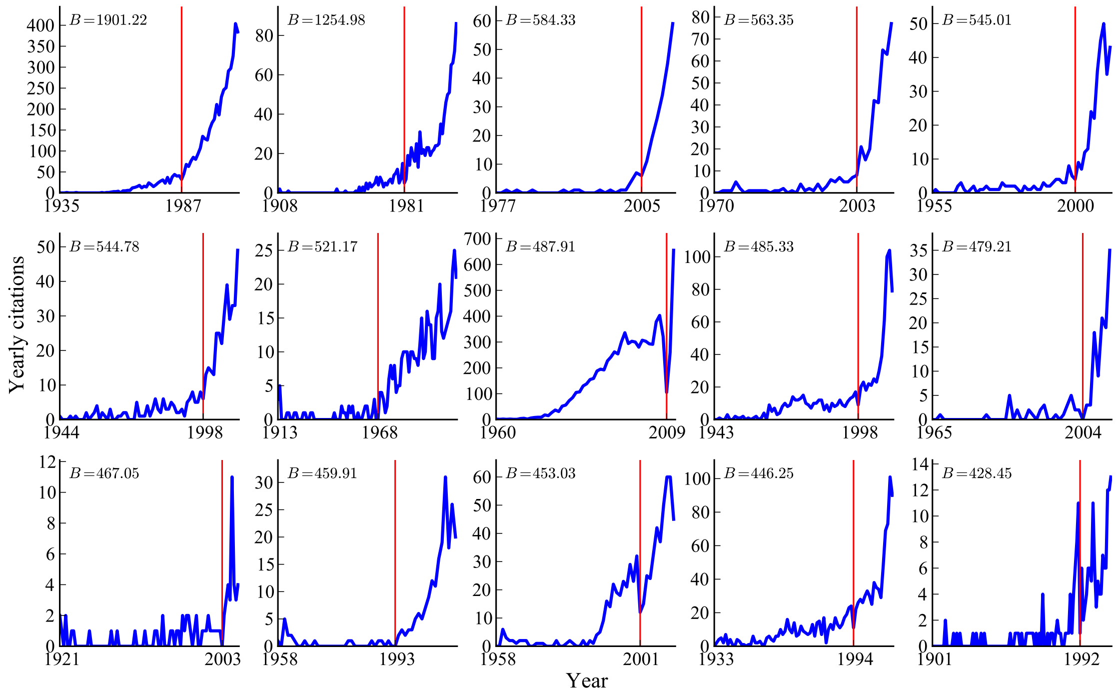

FIG. S7. (Blue) Citation histories, (Red) awakening years, and B values of top 15 Sleeping Beauties in Social Sciences and Humanities based on the B values in the WoS dataset. The ending year is 2011.

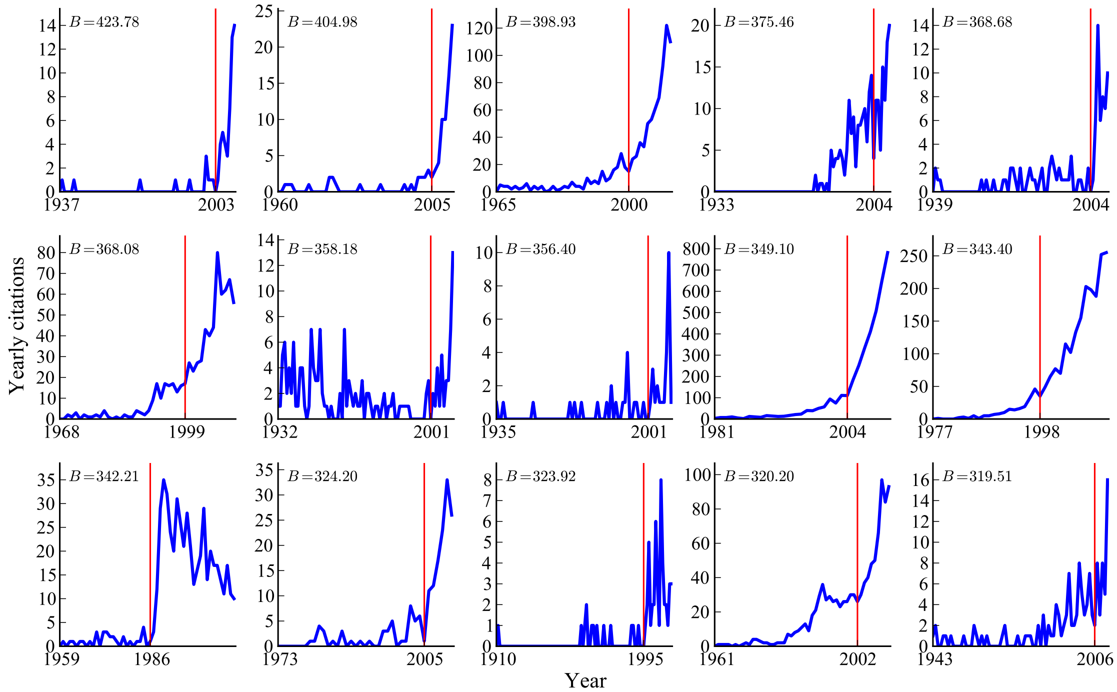

FIG. S8. (Blue) Citation histories, (Red) awakening years, and B values of 15 Sleeping Beauties ranked from 16^{th} to 30^{th} in Social Sciences and Humanities based on B values in the WoS dataset. The ending year is 2011.

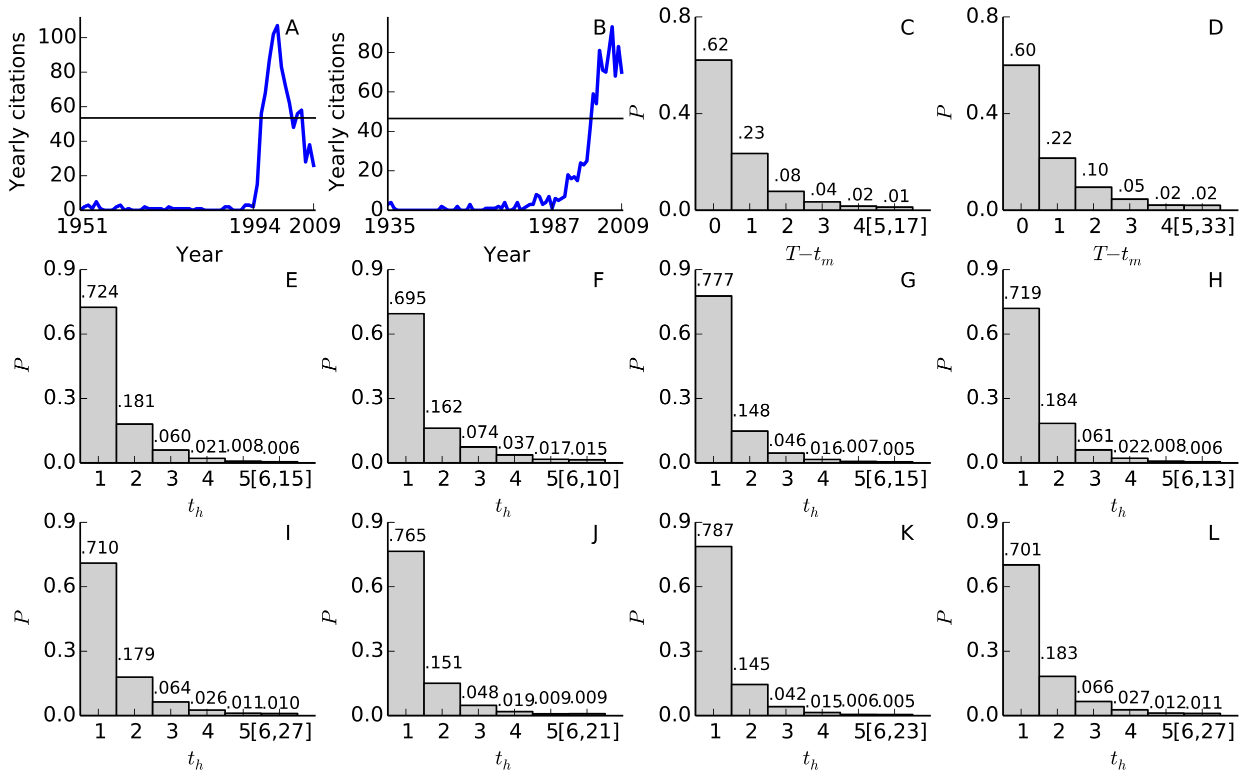

FIG. S9. Characterization of decreasing citation patterns of Sleeping Beauties. (*A*–*B*) Papers with positive beauty coefficient B are classified into two categories depending on whether or not their yearly citation counts have decreased to half of their maximum. (*C*–*D*) For papers belonging to the first class, we measure the length $T - t_m$ of the observation window at our disposal. $T = 2009$ for the APS and $T = 2011$ for the WoS are the last year covered by our datasets. t_m is instead the year when we observe the maximum number of yearly citations accumulated by an individual paper. The figures display the histograms of the quantity $T - t_m$ obtained for the APS (C) and WoS (D) dataset. (*E*–*H*) For papers that have experienced a fall in yearly citation counts at least below the half of their peak height c_m, we measure t_h, i.e., the number of years necessary to fall below the line $c_m/2$. We show that the distribution of t_h is insensible to the specific dataset considered, and to their beauty coefficient B. Panels F, G and H refer to the papers of the APS dataset ranked in the top 1%, top 1% to 10%, below 10%, respectively. Panels *I*–*L* show the same histograms as those of panels *E*–*H*, but for the WoS dataset.

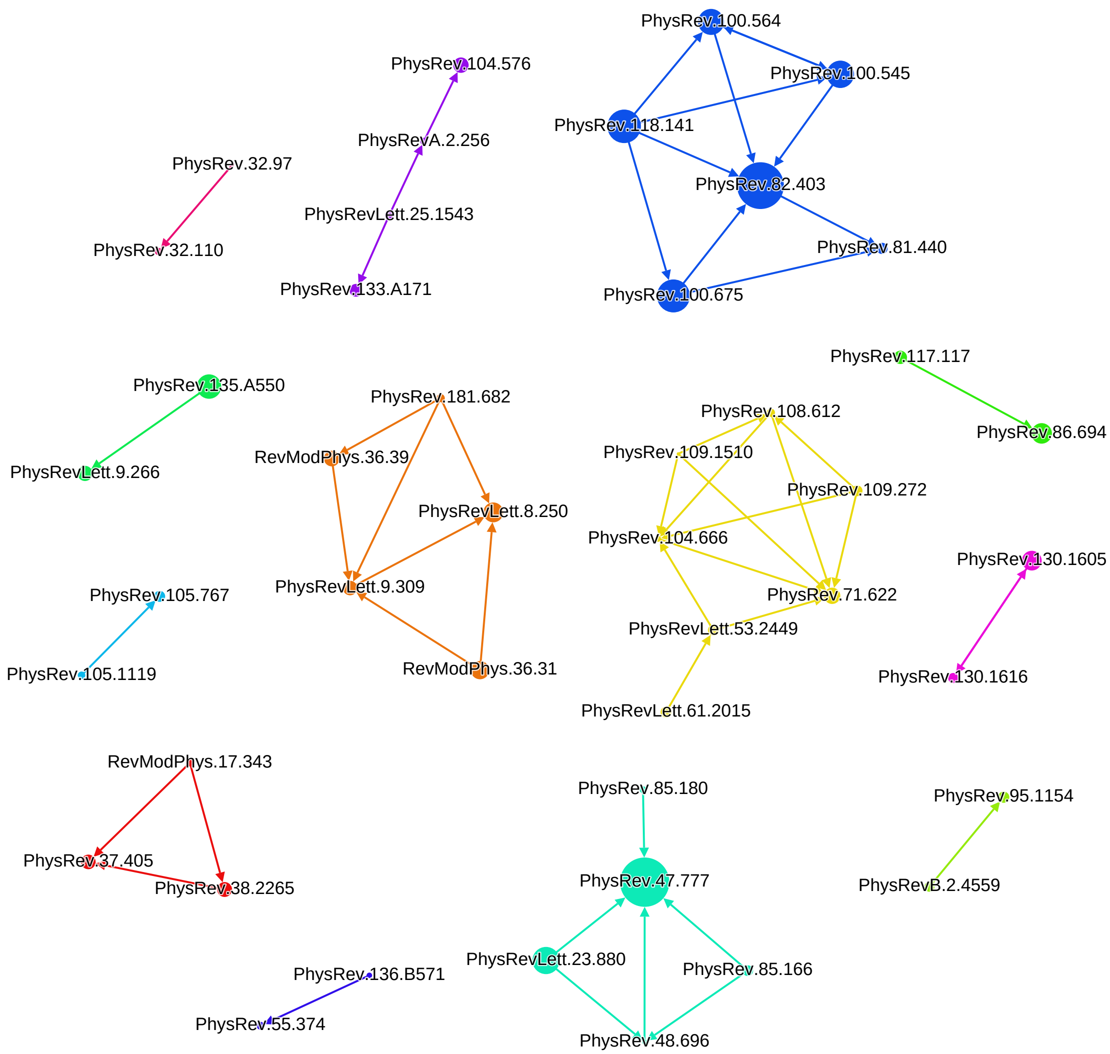

FIG. S10. The citation network of the 100 papers with highest B values in the APS dataset. Isolated nodes are omitted. The size of a node is based on its total number of citations.

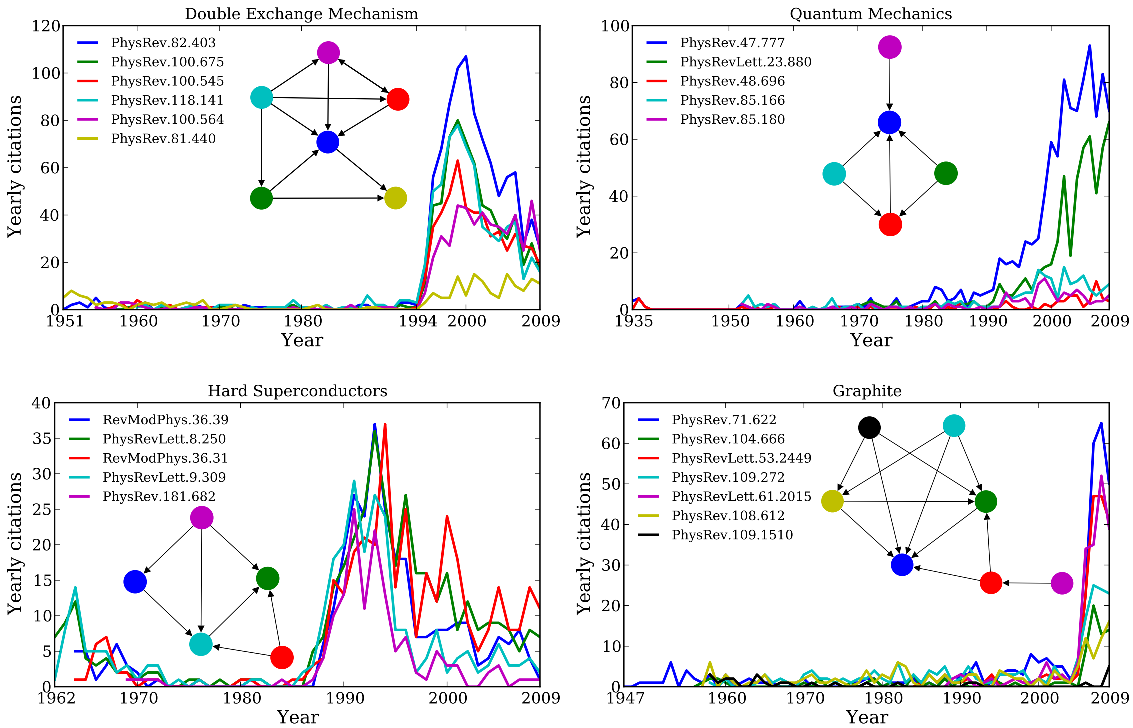

FIG. S11. The citation network reveals coarse topics of Sleeping Beauties. Papers belonging to the same group exhibit similar citation histories.

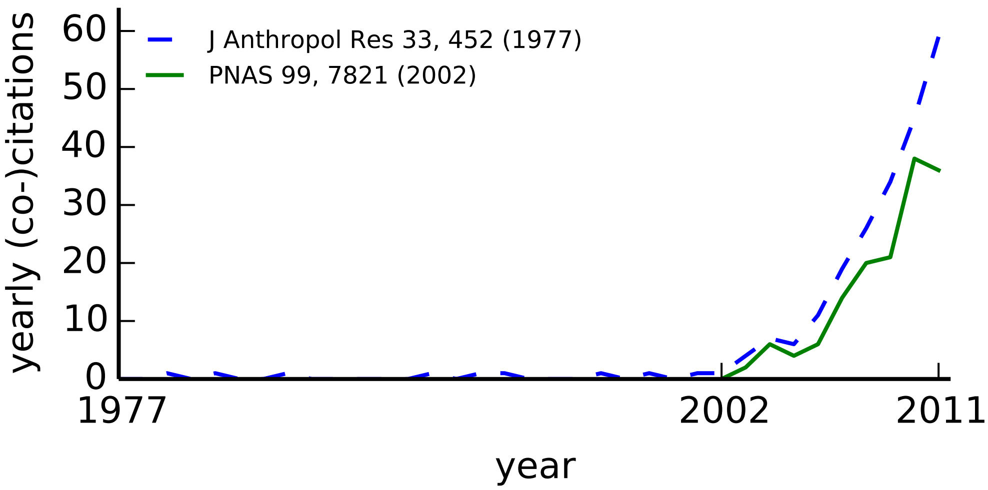

FIG. S12. Citation history of the paper J. Anthropol. Res. **33**, 452 (1977) [11]. The most co-cited paper is PNAS **99**, 7821 (2002) [6].

Subject category	Range of B
physics, multidisciplinary	[90.56, 5922.97]
chemistry, multidisciplinary	[90.57, 10769.06]
multidisciplinary sciences	[90.54, 3892.49]
mathematics	[90.62, 1215.38]
medicine, general & internal	[90.58, 1522.30]
physics, applied	[90.63, 3978.42]
surgery	[90.57, 799.65]
chemistry, inorganic & nuclear	[90.55, 1333.20]
statistics & probability	[90.56, 2736.18]
mechanics	[90.56, 3978.42]
biology	[90.68, 1247.13]
ecology	[90.60, 1792.29]
physics, condensed matter	[90.58, 3978.42]
biochemistry & molecular biology	[90.62, 839.22]
astronomy & astrophysics	[90.56, 984.81]
physics, atomic, molecular & chemical	[90.60, 774.23]
neurosciences	[90.59, 633.23]
materials science, multidisciplinary	[90.63, 3978.42]
plant sciences	[90.54, 1199.00]
engineering, chemical	[90.60, 2962.53]

TABLE S5. Threshold of B for each of the top 20 subject categories producing the top 0.1% SBs in the WoS dataset.

[1] P. Anderson and H. Hasegawa. Considerations on double exchange. *Phys. Rev.*, 100:675–681, Oct 1955.

[2] P. de Gennes. Effects of double exchange in magnetic crystals. *Phys. Rev.*, 118:141–154, Apr 1960.

[3] G. Dresselhaus. Spin-orbit coupling effects in zinc blende structures. *Phys. Rev.*, 100:580–586, Oct 1955.

[4] A. Einstein, B. Podolsky, and N. Rosen. Can quantum-mechanical description of physical reality be considered complete? *Phys. Rev.*, 47:777–780, May 1935.

[5] P. Fulde and R. Ferrell. Superconductivity in a strong spin-exchange field. *Phys. Rev.*, 135:A550–A563, Aug 1964.

[6] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, 2002.

[7] J. Goodenough. Theory of the role of covalence in the perovskite-type manganites $[\mathrm{La}, m(\mathrm{II})]\mathrm{Mno}_3$. *Phys. Rev.*, 100:564–573, Oct 1955.

[8] S. Redner. Citation statistics from 110 years of physical review. *Physics Today*, 58(6):49–54, 2005.

[9] P. Wallace. The band theory of graphite. *Phys. Rev.*, 71:622–634, May 1947.

[10] E. Wollan and W. Koehler. Neutron diffraction study of the magnetic properties of the series of perovskite-type compounds $[(1-x)\mathrm{La}, x\mathrm{Ca}]\mathrm{Mno}_3$. *Phys. Rev.*, 100:545–563, Oct 1955.

[11] W. W. Zachary. An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, 33(4):452–473, 1977.

[12] C. Zener. Interaction between the d-shells in the transition metals. ii. ferromagnetic compounds of manganese with perovskite structure. *Phys. Rev.*, 82:403–405, May 1951.