OPEN

Do scientists trace hot topics?

Tian Wei[1], Menghui Li[2,3], Chensheng Wu[1,4], Xiao-Yong Yan[1,5], Ying Fan[1], Zengru Di[1] & Jinshan Wu[1]

[1]School of Systems Science, Beijing Normal University, Beijing, 100875, P. R. China, [2]Temasek Laboratories, National University of Singapore, 117508, Singapore, [3]Beijing-Hong Kong-Singapore Joint Centre for Nonlinear & Complex Systems (Singapore), National University of Singapore - Kent Ridge, 119260, Singapore, [4]Beijing Institute of Science and Technology Intelligence, Beijing, 100044, P. R. China, [5]Centre for Complex Systems Research, Shijiazhuang Tiedao University, Shijiazhuang, 050043, P. R. China.

Do scientists follow hot topics in their scientific investigations? In this paper, by performing analysis to papers published in the American Physical Society (APS) Physical Review journals, it is found that papers are more likely to be attracted by hot fields, where the hotness of a field is measured by the number of papers belonging to the field. This indicates that scientists generally do follow hot topics. However, there are qualitative differences among scientists from various countries, among research works regarding different number of authors, different number of affiliations and different number of references. These observations could be valuable for policy makers when deciding research funding and also for individual researchers when searching for scientific projects.

T he phenomenon of "the rich get richer", which is also called preferential attachment in the field of complex networks[1], is quite common in many fields[2,3] (for example, see references cited in Table I); however, the scientific field is composed of scientists, a special group of people who focus on proposing, investigating and implementing original and creative ideas. Therefore, it is plausible that the "the rich get richer" phenomenon is less pronounced in the fields investigated by scientists than in other areas. Ideally scientists choose their fields of investigation according to their scientific interest and the scientific value of the investigated questions but not due to the hotness of the investigated fields. In this work, using published papers from the American Physical Society (APS) Physical Review journals beginning in 1976 and ending in 2009, we test whether the subject of a new paper is more likely to be in a hot field than in a relatively unknown field when the paper is published. We also compare scientists from different countries. Such comparisons could provide insightful and interesting information. In China, modern scientific development is still very young. It is believed among many scientists that there are many more Chinese scientists that are followers than original thinkers compared with many other countries. In this work, we offer direct empirical support for this hypothesis. Finally, we also determine if the degree of tracing hot fields differs for papers with different number of authors or affiliations and different number of references. Interestingly, it is found that scientists who collaborate with more authors or more affiliations tend to follow hotter fields than those who works with a few collaborators or affiliations. Moreover, papers with a small number of references, on average, are more interest-driven or value-driven, whereas papers with a large number of references are more hotness-driven. These empirical discoveries, particularly if it is also performed in other fields and for a larger periods of time, could provide valuable information for policy makers.

Results

Empirically, the phenomenon of preferential attachment has been found in many systems. We compile a list of typical systems, where their positive exponent α values are shown in Table I, indicating that new links are more likely to attach to nodes with larger degree. Later, we will compare our results on hotness tracing of newly published papers against other phenomena listed in this table.

Firstly, we examine the phenomenon of preferential attachment of papers in the PR-PACS data set (see *Methods* for details). In a log-log plot, Fig. 1 (a) displays the accumulated distribution function $\kappa(k)$ with respect to the size k of the field that a newly published paper belongs to. The positive exponents α indicate that new papers are more likely to focus on hot fields (larger sizes), i.e., new links preferentially attach to nodes (PACS) with larger sizes. Or to say, generally, scientists do publish more new papers in current hot fields. We obtain the exponents α by least-square fitting from $k = 1$ to $k = 300$ as the curves deviate from the straight line for large k due to low statistics. For different years t, all of $\kappa(k)$ follow power law, namely $\kappa(k) \sim k^{\alpha+1}$, but slightly different parameters α (as shown in Fig. 1 (b)). We also plot the distribution of fields' size $N(k)$ as an inset in Fig. 1 (a), which follows a highly skewed distribution.

Table I | A list of some values of α, the degree of preferential attachment in different networks collected from literature, showing also the number of nodes N (and M for bipartite networks), the number of links E. The first column is the name of the investigated database, and the discussed relationships are discussed within the brackets. Please note that APS-PACS, Google Earth and Digg are bipartite networks between individuals and groups while all other networks describe relations among individuals

Network	N (M)	E	α	Ref.
APS-PACS (papers belong to fields)	N:315,082 (M:5,472)	900,832	0.52–0.87	–
Digg (Membership of group)	N:212,635 (M:50)	1,185,167	0.69	5
Linkedin (friendship)	7,550,955	30,682,028	0.6	6
Sexual networks (sexual contact)	260–1220	–	0.32–0.80	4
Google Earth (Membership of group)	N:231,976 (M:54)	345,038	1.11	5
Medline (coauthorship)	1,648,660	–	1.04 ± 0.04	11
NYGI(coproduction)	10,000	700,000	1.20 ± 0.06	15
Bar (communication)	3,988	–	1.25 ± 0.13	16
Google(communication)	39,918	–	1.36 ± 0.14	16
Flickr (following)	584,207	3,554,130	1.0	6
aNobii (friendship, following)	86,800	697,910	1.0	17
Douban (following)	1,614,288	14,573,170	0.95	18
Wealink (friendship)	223,482	273,209	1.0	19
Citation (Citation)	1,736	83,252	0.95 ± 0.1	12

Compared with the preferential attachment phenomena in other fields listed in Table I, the exponents $\alpha = 0.52$–0.87 from our PR-PACS data are near the lower end of all the exponents. Out of all the other 13 entries in Table I, only sexual contact in sexual networks ($\alpha = 0.32$–0.80)[4], users attached to membership of groups of Digg ($\alpha = 0.69$)[5] and friendship relations in Linkedin ($\alpha = 0.6$)[6] are approximately at the same level of hotness tracing. It might be easy to "follow" a star member in a social networking website, such as Flickr ($\alpha = 1.0$)[6]; however, it might not be so easy to shift a research field, join a user group of a different product or change sexual partners towards hotter choices. Besides the difficulty in changing fields, it might also because that scientific works are more independent and more interest-driven or value-driven than other fields.

Secondly, we test whether the intensity of tracking hot topics of scientific research differs in different countries. Therefore, we classify the papers according to countries of the first author's affiliation, and calculate the absolute contribution ratios r^c of several major countries within PR-PACS data set. They are USA (33.07%), Germany (9.95%), Japan (6.94%) and China (3.73%). As seen from Fig. 2 (a), in the year 2008 the exponent ($\alpha = 1.06$) of P. R. China is larger than that of other countries, e.g., USA 0.74, Germany 0.86 and Japan 0.93. Moreover, as shown in the inset of Fig. 2 (a), the exponent α for China is generally larger than that of other countries for different years. These results indicate that the phenomenon of tracing hotness is

more severe among Chinese scholars. In addition, it is found that UK (3.80%) and France (5.55%) lie between USA and Germany, while Russia (1.89%) lies between Germany and China.

To provide a comparative study, we also calculate the relative contribution ratio $R^c(k)$ of papers from country c and belonging to the fields with size k (see *Methods* for the details). As seen from Fig. 2 (b), the relative contribution ratio $R^c(k)$ of Chinese scholars is smaller than 1 in cold fields (smaller k) but larger than 1 in hot fields (larger k), indicating that Chinese scholars make less contributions to cold fields than their average contribution but more to hot fields than their average. Meanwhile, the situation of USA is opposite to that of China. This difference also indicates that Chinese scientists are more keen to follow hot topics than United States scientists from another aspect. This agrees with our previous observations.

Considering the fact that scientific studies in China are still young, it is understandable that a large percentage of them are on hotness-driven fields rather than value-driven fields. The discovered order – the USA, Germany, Japan and China – of degree of hotness tracing makes sense intuitively. These results are more or less consistent with our intuitions. How different positions are related to scientific policies of that country, or even the culture and values of that country, although is definitely worth a further investigation, is outside the scope of the current study. We simply want to demonstrate the capability of the methods that are discussed above in analyzing
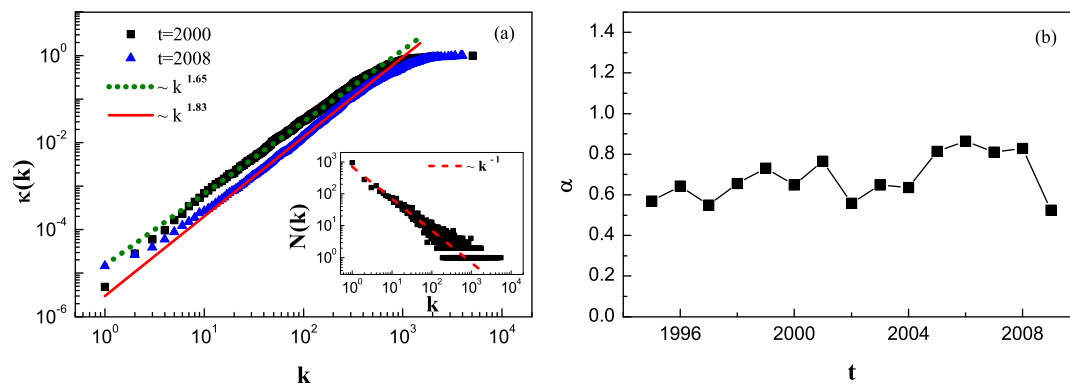

Figure 1 | Empirical preferential attachment to hot fields of new papers in the PR-PACS data set. (a) The cumulative probability functions $\kappa(k)$ in the years 2000 and 2008. Inset: the distribution function of the sizes of fields. (b) The exponents α for different years. We start our measurements from 1995, not 1976, the beginning year of the data, since the method requires a relatively large initial system. Notice $\alpha = 0.52$–0.87 is near the lower end of all the exponents collected in Table I. The 2009 exponent is relatively smaller compared to its previous several years for reason that we do not yet know. If not mentioned, Δt is one year in this paper. The straight lines are guide to the eye through this paper.
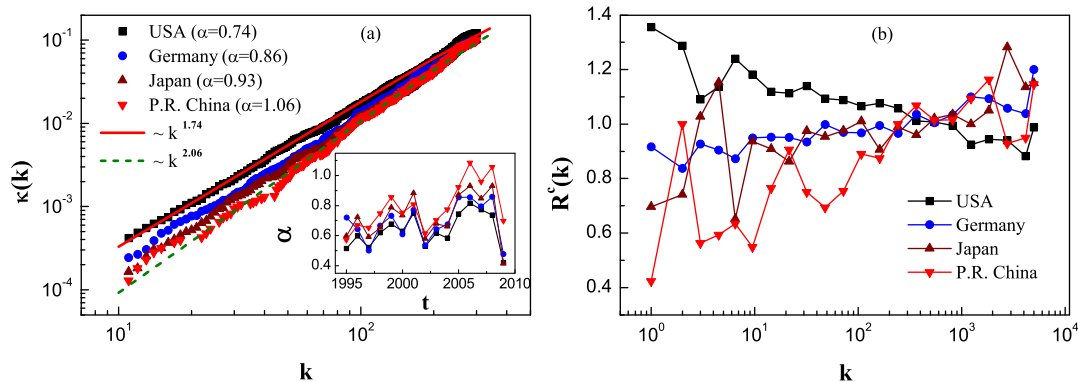
Figure 2 | Results for authors from the most-contributed countries in the PR-PACS data set. (a) The cumulative probability functions $\kappa(k)$ in the year 2008 for USA, Germany, Japan and P. R. China. The exponent of Chinese authors ($\alpha = 1.06$) is much larger than that of other countries. Inset: The exponents α for different years, where the exponents α of China always are larger than that of USA. (b) The relative contribution ratio $R^c(k)$ for USA, Germany, Japan and P. R. China.

publication records, and to present some basic discoveries using the methods in this work.

Next, we measure the influence of different number of authors and affiliations on the degree of tracing hot topics. Therefore, we classify the papers according to their number of authors and number of affiliations. It is argued in Ref. 7 that research works with many authors or many affiliations typically focus more on hot topics because it might exactly be the hotness of the paper subject that made collaboration attractive among the scientists and that a joint task team is generally more likely to focus on short-term projects rather than long-term projects. Here we make such an examination based on the PR-PACS data. We can see from Fig. 3 that overall, α, the degree of severity of tracing hot topics, increases with the number of authors and affiliations. These results provide empirical supports for the arguments in Ref. 7. In a sense, global collaboration is not necessarily a good strategy for high-quality, value-driven research topics as suggested in Ref. 7.

Moreover, it is also claimed that the size of teams and the share of multi-university collaboration grows rapidly in recent decades[8–10], and teams and multi-university collaborations tend to produce more highly cited papers[9,10]. It is believed that generally, the papers belonging to hot fields can easily acquire more citations than that of cold fields as there are more papers focusing on similar topics. Therefore, the fact discovered in[8–10] does not imply that multi-university collaboration and larger teams are making bigger and significant progress. It might more due to that fact that the fields they are working on is bigger. We check this hypothesis in this section to see if the papers with more authors and more affiliations are more likely to attach to hot fields. It is indeed so as shown in Fig. 3. From this view point, our results are consistent with the claims in Ref. 9,10.

Finally, we investigate the effect of the number of references on tracking hot topics. We classify the papers according to their number of references. It is obvious that the average number of references in papers today is much larger than that of early times. For earlier times, one can intuitively hypothesize that a pioneer paper or a paper of good quality typically cited less references. However, today, the number of references may or may not relate to how innovative the paper is. Here, we examine this hypothesis. As seen from Fig. 4, the exponents α, the degree of severity of tracing hot topics, increase with the number of references, which could indicate that papers with a larger number of references are more likely to be on hot topics. Notice that the absolute values of α for larger number of references in recent years are larger than that in earlier years, suggesting that scientists trace hot fields a bit more severely in recent years than in earlier times. Therefore, the hypothesis is reasonable overall.
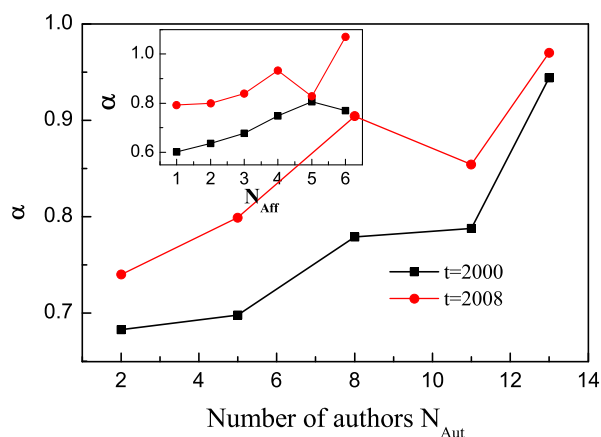

Figure 3 | The preferential attachment exponents α with respect to the number of authors N_{Aut} (the number of affiliations N_{Aff} in the inset) in the year 2000 and 2008. Note that every three numbers for authors are grouped together and labelled as the intermediate number, e.g., 1, 2 and 3 are grouped together and labelled as 2. Overall, the exponents α increase with number of authors and affiliations.
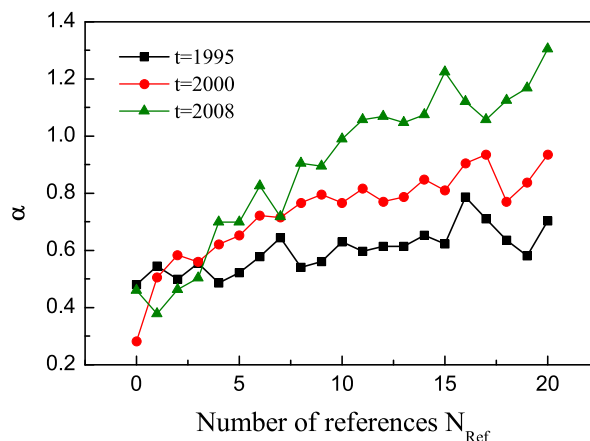

Figure 4 | The preferential attachment exponents α with respect to the number of references N_{Ref}. Clearly, the exponents α increase with the number of references. The absolute value of α is a bit larger in recent years than in early years for larger number of references.

Discussion

We have found that in the tested fields of science for papers published by the APS physical review journals, hot fields attract more newly published papers; however, scientific works are not as hotness-driven compared with other fields. Among the major countries, this phenomenon of tracing hotness occurs more in China compared with other countries, which could be related to the fact that China is still a developing country in terms of scientific research. We also found that papers with more authors, more affiliations and more references were on subjects that were more hotness-driven. This observation could potentially be valuable to policy makers that fund scientific projects.

Here only data of physics publications were analyzed. A cross-field comparison would be interesting, particularly if one can do similar studies on math and social science, such as economics. Similar studies can be applied on producing, selling or buying products. That is to investigate when a product is manufactured, sold or bought, how often is it related to the number of the product that has been manufactured, sold or bought already. Such studies could be valuable for marketing research.

It might also be interesting to determine how often outstanding/important works when they are first published are in hot fields. Fields gradually or suddenly become hot after major breakthroughs are presented in a few pioneer papers, which later on, might be awarded or honoured. Using records of awards such as the Nobel prize, the Fields Medal, the Turing Award etc., it would be possible to perform a statistical analysis of papers with major breakthroughs. All the award-winning papers could be collected, where the same analysis to determine α can be performed, which could then be compared with the overall α.

The above investigation can be performed, not only at the macro level of counties of authorship but also at the mesoscopic level of affiliations of authorship. In this way, one might be able to compare tradition, culture and research quality at various levels of academic units. In principle, one could also collect all published papers of one author and treat all those papers as a subset and study the preferential attachment phenomenon of those papers if the size of such collection is large enough. It would then be possible to use it to measure the creativity and originality of a researcher.

Methods

Data description and notations. Our data set is a collection of all papers published by the American Physical Society (APS) Physical Review journals from 1976 to 2009. Each paper is recorded as a data entry, which includes its title, date of publication, classification number according to the AIP-Physics and Astronomy Classification Scheme (PACS), author(s), affiliation(s) and reference(s) to other papers within the data set. The entire data set contains more than 320, 000 entries, including a variety of article types, such as article, brief report, rapid communication, comment, reply, erratum, essay, announcement, editorial, announcement and so on. Here, we will only consider those research papers, e.g., article, brief report and rapid communication, with PACS numbers and refer to it as the PR-PACS data set. At last, we have $N = 315,082$ entries, which each entry, i.e. each paper, is denoted as i, and $M = 5,472$ PACS, which each entry, i.e. each PACS number, is denoted as λ.

We use PACS, the established hierarchical classification systems of physics, to identify the fields within the physics field. The investigated system of APS-PACS is a bipartite network, where there are two kinds of nodes, i.e., papers and PACS, and one kind of links, i.e., relation between paper and PACS, which is determined by the researchers. Mathematically, we can use one matrix to characterize the relation between paper and PACS. A is an $N \times M$ adjacency matrix, with element $a_{i\lambda} = 1$ if paper i belongs to PACS λ, or to say, PACS λ appears in paper i, otherwise 0. Generally, one paper can belong to several fields (have several PACS), therefore there are a total of 900,832 links between papers and PACS, i.e., $E = \sum_{j,\lambda} a_{i\lambda}$. The size of a field, $K_\lambda = \sum_i a_{i\lambda}$, is defined as the number of papers that belong to it, describing the popularity (hotness) of the fields. Then, we can calculate the number $N(k)$ of fields with size k. Based on PR-PACS data, our main concern is to observe whether new papers would like to attach to hot fields (with larger size k), i.e., whether new links preferentially attach to the nodes with larger degree.

Measuring preferential attachment. Here, we explain our method for this statistical analysis for testing preferential attachment on temporal data set. The basic idea is to investigate whether new links are likely to attach to nodes with larger degree (size).

We calculate the empirical value of the relative probability $T(k)$ that a new paper published within a short period Δt connects to a field which has a size of k before the time t[11] as follows. Since the corresponding time-dependent absolute probability $P_k(t)$ that a new paper published in a field with size k is proportional to $T(k)n_k(t)/N(t)$, where $n_k(t)$ is the number of fields with size k and $N(t)$ is the number of fields immediately before time t, then $T(k)$ can be estimated by making a histogram of the sizes k of the fields to which each paper is added within the time period Δt in which each sample is weighted by a factor of $\frac{N(t)}{n_k(t)}$,

$$T(k) = \sum_{i,\lambda}^{k_\lambda(t) = k} \frac{a_{i\lambda} N(t)}{n_k(t)} \tag{1}$$

where $k_\lambda(t) = k$ means that the field, to which the papers published within the period Δt belong, has size k at time t. We now have the empirical curve $T(k)$ from the above statistical analysis. In order to conveniently compare $T(k)$ with different time t, $T(k)$ is normalized as $T'(k) = \frac{T(k)}{\sum_{k'} T(k')}$[12,13].

The preferential attachment hypothesis states that the rate $T'(k)$ with which a node with k links acquires new links is a monotonically increasing function of k[1], namely

$$T'(k) = \frac{k_i^\alpha}{\sum_j k_j^\alpha} = C(t) k_i^\alpha \tag{2}$$

For BA model $\alpha = 1$[1]. To obtain a smooth curve from noisy data, we take the cumulative function form instead of $T'(k)$:

$$\kappa(k) = \int_0^k T'(k) \, dk' \tag{3}$$

Thus, $\kappa(k)$ should be proportional to $k^{\alpha+1}$. We can now fit the empirical curve from the previous statistical analysis and then compare it against this hypothesized curve of preferential attachment. This is the general procedure of all the analysis presented in this work. This method has been extensively used to measure the preferential attachment phenomenon in different kinds of networks[4–6,11–19].

To test the preferential attachment of scientific research differs in different countries, we separated the entire data set according to countries of the first author's affiliation and then perform a comparison among the most contributed countries (USA, Germany, Japan, China). With this separated data set, we perform the examination of preferential attachment only counting the papers from authors in country c as

$$T^c(k) = \sum_{i,\lambda}^{k_\lambda(t) = k, aff_i = c} \frac{a_{i\lambda} N(t)}{n_k(t)}. \tag{4}$$

Here $aff_i = c$ means the principle affiliation of this paper i is in country c. In counting k and $n_k(t)$, we included papers from all countries, meaning that scientists from all countries face the temptation of tracing the same overall hotness in the entire PR-PACS data set. Similarly, besides countries, the above calculation can be applied to any features of papers, such as different number of authors, affiliations and references.

Measuring relative contribution ratio $R^c(k)$. Absolute contribution from a country is measured simply by a percentage of published papers from that country out of the total number of published papers,

$$r^c = \frac{\sum_k m_k^c}{\sum_k m_k}, \tag{5}$$

where m_k (m_k^c) is the number of papers (from country c) belonging to fields with size k. Here we present a more detailed breakdown of this absolute contribution by looking at each individual field what is the percentage of papers from that country out of all papers in that field, and then normalized by the absolute contribution of that country,

$$R^c(k) = \frac{m_k^c}{m_k} \frac{1}{r^c}. \tag{6}$$

This is a static measure, so it is easy to perform. In a sense it also describes how often scientists in that country are pursuing hot fields.

1. Barabási, A. L. & Albert, R. Emergence of scaling in random networks. *Science.* **286**, 509–512 (1999).
2. Simon, H. A. On a class of skew distribution functions. *Biometrika.* **42**, 425–440 (1955).
3. Price, D. D. S. A general theory of bibliometric and other cumulative advantage processes. *J. Am. Soc. Inf. sci.* **27**, 292–306 (1976).
4. De Blasio, B. F., Svensson, A. & Liljeros, F. Preferential attachment in sexual networks. *Proc. Natl. Acad. Sci. USA.* **104**, 10762–10767 (2007).
5. Shi, X., Zhu, J., Cai, R. & Zhang, L. User grouping behavior in online forums. *KDD '09.* 777–786 (2009).

6. Leskovec, J., Backstrom, L., Kumar, R. & Tomkins, A. Microscopic evolution of social networks. *KDD '08.* 462–470 (2008).
7. Adams, J. Collaborations: The rise of research networks. *Nature.* 490, 335–336 (2012).
8. Adams, J. D., Black, G. C., Clemmons, J. R. & Stephan, P. E. Scientific teams and institutional collaborations: Evidence from U.S. universities, 1981–1999. *Res. Policy.* 34, 259–285 (2005).
9. Wuchty, S., Jones, B. F. & Uzzi, B. The increasing dominance of teams in production of knowledge. *Science.* 316, 1036–1039 (2007).
10. Jones, B. F., Wuchty, S. & Uzzi, B. Multi-university research teams: shifting impact, geography, and stratification in science. *Science.* 322, 1259–1262 (2008).
11. Newman, M. E. J. Clustering and preferential attachment in growing networks. *Phys. Rev. E.* 64, 025102(R) (2001).
12. Jeong, H., Néda, Z. & Barabási, A. L. Measuring preferential attachment in evolving networks. *Europhys. Lett.* 61, 567–572 (2003).
13. Li, M., Gao, L., Fan, Y., Wu, J. & Di, Z. Emergence of global preferential attachment from local interaction. *New. J. Phys.* 12, 043029 (2010).
14. Szell, M. & Thurner, S. Measuring social dynamics in a massive multiplayer online game. *Social Networks.* 32, 313–329 (2010).
15. Saavedra, S., Reed-Tsochas, F. & Uzzi, B. Asymmetric disassembly and robustness in declining networks. *Proc. Natl. Acad. Sci. USA.* 105, 16466–16471 (2008).
16. Eom, Y. H., Jeon, C., Jeong, H. & Kahng, B. Evolution of weighted scale-free networks in empirical data. *Phys. Rev. E.* 77, 056105 (2008).
17. Aiello, L. M., Barrat, A., Cattuto, C., Ruffo, G. & Schifanella, R. Link creation and profile alignment in the aNobii social network. *SocialCom '10: Proceedings of the Second IEEE International Conference on Social Computing.* 249–256 (2012).
18. Zhao, J. Z., Lui, J. C. S., Towsley, D., Guan, X. H. & Zhou, Y. D. Empirical analysis of the evolution of follower network: a case study on Douban. *Computer Communications Workshops (INFOCOM WKSHPS), 2011 IEEE Conference.* 924–929 (2011).
19. Hu, H. B. & Wang, X. F. Disassortative mixing in online social networks. *EPL.* 86, 18003 (2009).

Acknowledgements

Author contributions
J.W., Z.D. and Y.F. designed the research. T.W., M.L., X.Y. and C.W. analyzed the data. T.W. and J. Wu participated in the writing of the manuscript.

Additional information

5