Eur. Phys. J. B (2013)
DOI: 10.1140/epjb/e2013-40132-2

THE EUROPEAN PHYSICAL JOURNAL B

Regular Article

The decentralized flow structure of clickstreams on the web*

Lingfei Wu[1] and Jiang Zhang[2,a]

[1] Department of Media and Communication, City University of Hong Kong, Hong Kong, P.R. China
[2] School of Management, Beijing Normal University, Beijing 100875, P.R. China

Received 17 February 2013 / Received in final form 19 April 2013
Published online (Inserted Later) – © EDP Sciences, Società Italiana di Fisica, Springer-Verlag 2013

Abstract. The browsing behavior of massive web users forms a flow network transporting user' collective attention between websites. By analyzing the circulation of the collective attention we discover the scaling relationship between the impact of sites and their traffic. We construct three clickstreams networks, whose nodes were websites and edges were formed by the users' switching between sites. The impact of site i, C_i, is measured by the clickstreams controlled by this site in the circulation of clickstreams. We find that C_i scales sublinearly with A_i, the traffic of site i. Specifically, there existed a relationship $C_i \sim {A_i}^{\gamma}(\gamma < 1)$, which implies the decentralized structure of the clickstream circulation.

1 Introduction

The explosive growth of the world wide web in the past two decades presents an urgent challenge for developing a quantitative, predictive theory of the interaction between the web and users. While previous studies analyzing the hyperlinks [1–5] and individual browsing records [6–9] provide insight for understanding surfing behavior, they have limitations restricted by the data used. Firstly, hyperlinks are too simple to represent the rich interactions between sites. From bookmarks and default home pages to historical viewing records, there are many different ways in which clickstreams are generated between sites of no hyperlink connections [10]. Secondly, while individual surfing behavior has been extensively investigated [7–9,11], there is still a lack of research studying collective browsing behavior from a network perspective [12]. Last but not least, if we want to understand the long-range, complex interactions between sites, the investigation on the local statistics of sites, such as hyperlink degree [2] or traffic [7,8], is not enough. Instead, we should probe into the transportation of traffic between sites, that is, the flow of clickstreams [12,17]. As an illustration, in this work we analyze the circulation of clickstreams among the top sites in the world.

In literature, there are generally two different opinions concerning clickstream dynamics. The "rich-get-richer" paradigm suggests that user navigation enlarges the inequality of traffic among sites [13–15]. On the contrary, the "egalitarian" paradigm argues that user navigation actually makes the web a level-play place where new sites have a greater chance of acquiring popularity [16]. To examine these two assumptions, we collect data from Alexa (`www.alexa.com`) and construct website-level clickstream networks [12,16,17]. In these networks, the nodes are websites and the edges are formed by the users' switching between websites. We use C_i to denote the impact of the ith site on other sites in clickstream circulation and examine its correlation with A_i, the traffic of the site. It turns out that C_i scales sublinearly with A_i as $C_i \sim {A_i}^{\gamma}(\gamma < 1)$. We suggested that this pattern reflects the decentralized structure of the studied clickstream networks. That is, compared to large sites, small sites had a disproportionately larger impact in the circulation of clickstreams. Therefore, our finding supports the assumption of the "egalitarian" paradigm.

The presented approach of clickstream network analysis is not only interesting at its own right, but also provides a new way to investigate online activities. For example, traditional studies on news diffusion focused on the diffusion of news among users [18,19], but from the perspective of clickstream analysis, we can also understand it in a "reversed" way, that is, the allocation of users' attention among news [20]. As a consequence, the rise and decay of news reflects the competition among them for users' collective attention [20]. We can also easily extend this analysis of news to tags [21], videos [22] or any other type of information resources.

2 Materials and methods

2.1 Data collection

We select three lists of top 1000 sites worldwide as seed sites. Two of them are collected from Google[1] and the rest one is collected from Alexa (please refer to supplementary material for the detailed information of these lists).

* Supplementary material in the form of one pdf file available from the journal web page at `http://dx.doi.org/10.1140/epjb/e2013-40132-2`

[a] e-mail: `zhangjiang@bnu.edu.cn`

[1] `http://www.google.com/adplanner/static/top1000/`

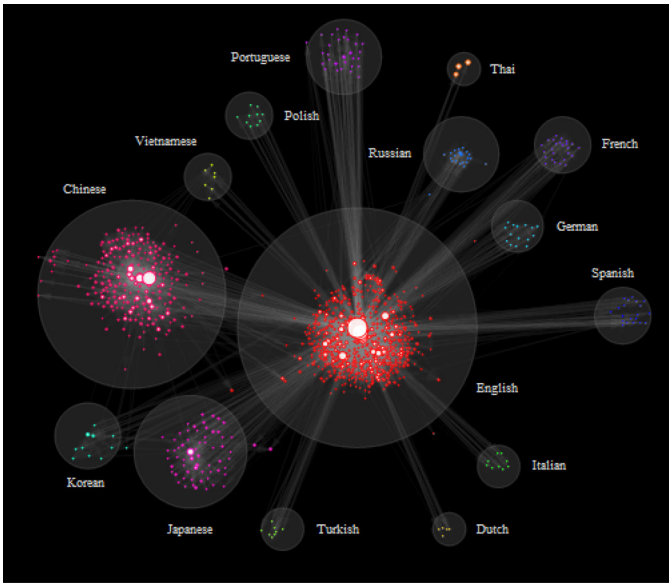

Fig. 1. The visualization of $w2$. Solid circles represent websites and edges show the clickstreams. The size of circles is proportional to the logarithmic value of their traffic. Websites of the same language are placed together and assigned the same color. The layout of the network is obtained by applying force-based algorithm [24] twice, firstly in the (language) community-level and secondly in the website-level within each community.

The clickstreams between the seed sites are downloaded from Alexa. From the downloaded data we construct three clickstream networks (which are called $w1$, $w2$, and $w3$ hereafter), in which a directed, weighted edge w_{ij} indicated the daily percentage of the global web users switching from site i to site j. As Alexa only reports the top ten inbound and outbound clickstreams for each site, our dataset does not necessarily include all the clickstreams between the studied sites. However, this does not mean that our sampling of clickstreams is lack of representativeness. Popular sites such as Google or Facebook may receive clickstreams from many other sites, therefore the degree distributions of the constructed networks are still long-tail. Actually, we can regard the studied networks as the "backbone" of the clickstreams on the entire web [23]. The detailed information of the three networks, including the degree distribution, is given in the supplementary material. In Figure 1, we plot $w2$ as an example. For visually appealing we put sites of the same language together and render them in the same color.

2.2 The definition of A_i, C_i, and γ

An example clickstream network (Fig. 2a) is given to illustrate the calculation of A_i, C_i, and γ. We balance the network by adding two artificial nodes, "source" and "sink", to make sure that at each node the sum of inbound and outbound streams are equal [25]. Suppose node i is imbalanced, that means $w_{ij} \equiv \sum_{j=1}^{n} f_{ji} - \sum_{j=1}^{n} f_{ij} \neq 0$. If $w_{ij} > 0$, which means the influx to i is larger than the

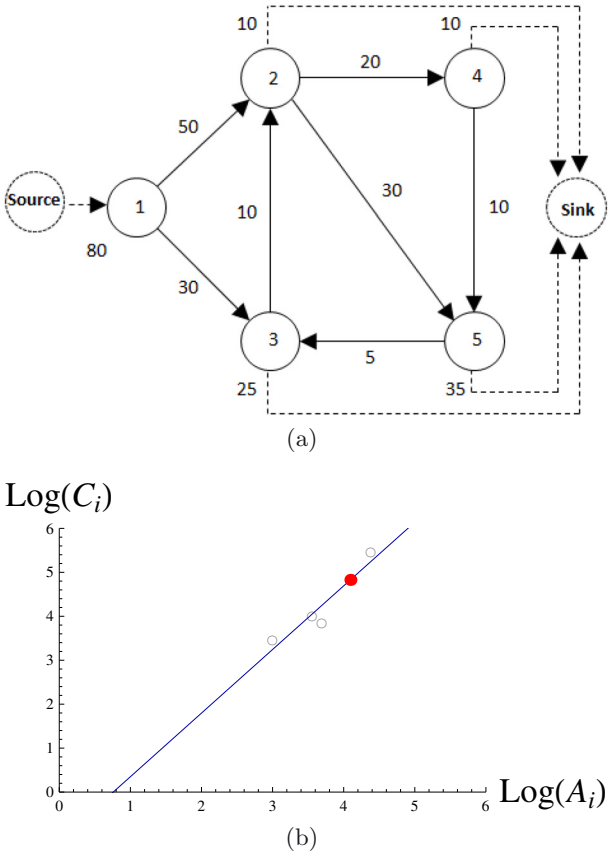

Fig. 2. (a) An example clickstream network and (b) the fittings of γ in the example network. The red node denotes the values of A_2 and C_2.

out flows from i, then we add a new edge from i to the sink $(n+1)$ with flux $|w_{ij}|$. Otherwise, if $w_{ij} < 0$, then a new edge from the source (0) to i with flux $|w_{ij}|$ is added. After that, we derive the matrix form of the balanced network (F') and normalize this matrix by row to obtain the transition matrix M (Eq. (1)), whose element m_{ij} denotes users' switch probability from site i to site j. Note that there are $n+1$ rows (columns) in M for we remove "sink" by remain "source" in the network in order to make equation (4) true.

$$m_{ij} = \frac{f'_{ij}}{\sum_{k=1}^{n+1} f'_{ik}}, \qquad \forall i, j = 0, 1, \ldots, n. \tag{1}$$

We define A_i and C_i as follows:

$$A_i = \sum_{k=1}^{n+1} f'_{ik}, \qquad \forall i = 1, 2, \ldots, n, \tag{2}$$

$$C_i = G_i \sum_{k=1}^{n} u_{ik}, \qquad \forall i = 1, 2, \ldots, n. \tag{3}$$

In equation (3), u_{ij} is the element of

$$U = \frac{1}{I - M} = I + M + M^2 + \cdots + M^{\infty} \tag{4}$$

F

$$\begin{pmatrix} 0 & 50 & 30 & 0 & 0 \\ 0 & 0 & 0 & 20 & 30 \\ 0 & 10 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 10 \\ 0 & 0 & 5 & 0 & 0 \end{pmatrix}$$

F'

$$\begin{pmatrix} 0 & 80 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 50 & 30 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 20 & 30 & 10 \\ 0 & 0 & 10 & 0 & 0 & 0 & 25 \\ 0 & 0 & 0 & 0 & 0 & 10 & 10 \\ 0 & 0 & 0 & 5 & 0 & 0 & 35 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

M

$$\begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{5}{8} & \frac{3}{8} & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{3} & \frac{1}{2} \\ 0 & 0 & \frac{2}{7} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{2} \\ 0 & 0 & 0 & \frac{1}{8} & 0 & 0 \end{pmatrix}$$

U

$$\begin{pmatrix} 1 & 1 & \frac{3}{4} & \frac{7}{16} & \frac{1}{4} & \frac{1}{2} \\ 0 & 1 & \frac{3}{4} & \frac{7}{16} & \frac{1}{4} & \frac{1}{2} \\ 0 & 0 & \frac{42}{41} & \frac{7}{82} & \frac{14}{41} & \frac{28}{41} \\ 0 & 0 & \frac{12}{41} & \frac{42}{41} & \frac{4}{41} & \frac{8}{41} \\ 0 & 0 & \frac{3}{164} & \frac{21}{328} & \frac{165}{164} & \frac{21}{41} \\ 0 & 0 & \frac{3}{82} & \frac{21}{164} & \frac{1}{82} & \frac{42}{41} \end{pmatrix}$$

Fig. 3. Summary of the steps in deriving matrix U.

Table 1. The statistics of three studied clickstream networks.

Network	N_{sites}	N_{edges}	Daily clickstreams	γ	R^2 of γ
w1	979	11 906	5.45×10^9	0.95	0.98
w2	956	11 529	1.38×10^{10}	0.92	0.95
w3	1189	17 061	6.06×10^9	0.96	0.99

Note: the daily clickstreams is obtained by summing up the number of unique users over all edges in a clickstream network.

and G_i is defined as:

$$G_i = \frac{\sum_{j=1}^{n} f'_{0j} u_{ji}}{u_{ii}}, \tag{5}$$

where f'_{0j} is the flow from "source" to j. Putting together equations (4) and (5) G_i is total flow transported from "source" to i [26,27] along all possible pathes (except the looping flow on i). Basing on the data of the example network, Figure 3 gives a summary of the calculations, which enables us to examine the scaling relationship:

$$C_i \sim {A_i}^{\gamma}, \tag{6}$$

in which A_i is the traffic of site i and C_i reflects the circulated clickstreams controlled by i in the network. As suggested in references [25,28,29], we can examine the scaling relationship between A_i and C_i. For example, Figure 1b plots $\log(C_i)$ against $\log(A_i)$ in the example network, in which the value of γ is estimated to be 1.45.

3 Results

3.1 The scaling of clickstreams

As shown in Figure 4, we find a scaling relationship $C_i \sim {A_i}^{\gamma}$ that is ubiquitous across the three studied networks. The value of γ is estimated to be in the range of $0.92 \sim 0.96$ (Tab. 1).

In Figure 4, we ignore the differences between users in investigating the scaling property of the clickstream networks. However, this assumption is naive considering the different preferences of users in navigation [6,7]. Therefore, in the following part we try to control the linguistic variance, which may be one of the most significant differences between users [16,30], in investigating the scaling property. In particular, we divide the clickstream networks into language-based website communities and then observe the scaling pattern across these communities.

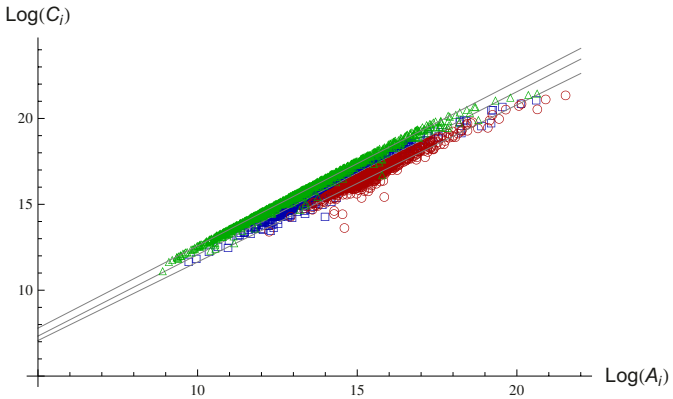

Fig. 4. The scaling relationship between A_i and C_i in the three clickstream networks. The data points from three networks are plotted in different colors and styles: blue squares for $w1$, red circles for $w2$, and green triangles for $w3$. The values of γ (black line) were 0.95, 0.92, and 0.96, respectively. Please refer to Table 1 for more information concerning the fitting of the scaling relationship.

Using the AlchemyAPI (`http://www.alchemyapi.com/`), we detect 16 language communities from $w1$, 17 from $w2$, and 50 from $w3$. In Table 2, we present the results of $w2$ as an example (the results of the rest two networks are given in the supplementary material). The communities given by Table 2, are less than those shown in Figure 1, for several communities are too small to support an estimation of γ. As suggested by Table 2 and Figure 5, in most of the communities there exists the relationship $C_i \sim {A_i}^{\gamma} (\gamma < 1)$, and the value of γ seems to be invariant of community size. It means that these communities share the common decentralized structure with the networks they belong to. This finding also implies that, the linguistic variance between users does not affect the universal scaling regularity in collective navigation [30].

3.2 The meaning of γ

The scaling relationship between C_i and A_i is interesting at its own interest, but it is particularly inspiring to

Table 2. The scaling exponent across language communities in $w2$.

Community	N_{sites}	N_{edges}	Daily clickstreams	γ	R^2 of γ
English	516	6188	8.64×10^9	0.94	0.94
Chinese	214	2130	3.18×10^9	0.94	0.77
Japanese	63	481	4.83×10^8	0.86	0.88
Portuguese	31	115	4.48×10^7	0.91	0.83
French	25	57	1.12×10^7	0.84	0.57
Russian	21	94	8.50×10^7	0.97	0.94
German	15	64	1.78×10^7	0.91	0.76
Korean	11	53	6.11×10^7	0.98	0.84
Polish	10	43	1.72×10^7	1.05	0.91
Vietnamese	7	25	8.70×10^6	0.86	0.61
Thai	3	6	1.67×10^6	0.31	0.71

Note. The daily clickstreams is derived by summing up the number of unique users over all edges in the clickstream network.

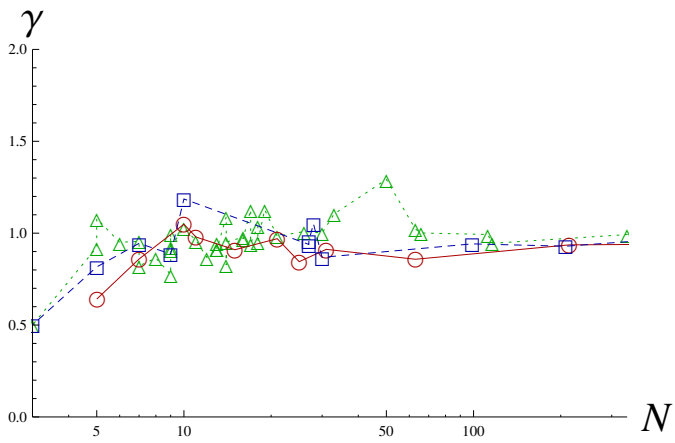

Fig. 5. The change of γ with community size N. Each data point corresponds to a language community. The data points of the three networks are plotted in blue squares ($w1$), red circles ($w2$), and green triangles ($w3$), respectively.

consider the possible interpretations of γ. If we treat C_i as an indicator of the total (both direct and indirect) impact of site i on the rest of sites in clickstream circulation [25,31], then γ measures the average increase of impact with traffic. We can interpret γ as the level at which large sites dominate the circulation of clickstreams [32]. For example, suppose we have two clickstream networks of the same traffic distribution but are different in γ, $A_i = \{1, 2, 3, 4, 5\}$, $\gamma' = 1/2$, and $\gamma'' = 2$. We can derive that $C_i' = \{1, 1.4, 1.7, 2, 2.2\}$ and $C_i'' = \{1, 4, 9, 16, 25\}$. The impact of the largest node is 2.2 in the former network, meaning that it controls $(2.2/(1 + 1.4 + 1.7 + 2.2)) \approx 27\%$ circulated clickstreams. However, we can calculate that in the latter network the largest node controls 45% circulated clickstreams. Therefore, the latter network is more heavily dominated by large sites. To conclude, $\gamma < 1$ implies a decentralized flow structure, whereas $\gamma > 1$ is the signature of a centralized flow structure. Our finding of the $\gamma < 1$ in empirical clickstream networks uncovers the decentralizing nature of collective attention, which is consistent with the finding of reference [16].

3.3 The robustness of the scaling pattern against network permutations

The necessity of the work presented in this section is twofold. Firstly, to overcome the limitations of the currently used data sets. As Alexa only provides the top ten inbound and outbound clickstreams for each site, we have to ignore the rest, smaller clickstreams in the data analysis. If the discussed scaling pattern is sensitive to the missing of the clickstreams, our conclusion would be biased. Secondly, by testing the robustness of the scaling pattern against network permutations we obtain insight into the mechanism responsible for the observed pattern.

We investigated the robustness of the scaling relationship against two types of network permutations, the selective removal of small clickstreams and the random shuffling of connections. In both analysis, we used four statistics to describe the scaling pattern, including γ, R^2, ρ, and D. γ and R^2 are the fitted parameter and the explained variance of the OLS regression of $\log(C_i)$ on $\log(A_i)$, respectively. ρ is the Pearson correlation coefficient between $\log(C_i/A_i)$ and $\log(A_i)$ [29] and D is the Kolmogorov-Smirnov distance between C_i and ${A_i}^{\gamma}$. As γ is close to 1, it is difficult to determine whether the observe scaling relationship is a trivial, linear dependence or a significant, non-linear pattern. But by calculating ρ, we are able to examine the non-linear nature of data, because in $\log(C_i/A_i)$ we have removed the linear trend. In particular, $\rho \approx 0$ if there is only linear relationship between C_i and A_i, and $\rho \ll 0$ if C_i scales with A_i sublinearly. D is often used, particulary, in cases concerning skewed distributions, to determine whether two data sets are from the same population (the KS test) [33,34]. We calculate D and compare it with 0.035, which is the expected value of D under a confidence level equals 0.1 [34] and a sample size equals 1200 [33]. If $D < 0.035$, the discussed scaling relationship is validated.

In the first analysis, we gradually removed small clickstreams from the networks and observed the change of the statistics [23]. We defined $0 \leq \alpha < 1$ as the fraction of the kept clickstreams for each site and remove all other clickstreams. Figure 6 shows that the scaling pattern is robust against the removal of edges. The values of γ and R^2 do not fluctuate a lot as long as more than 30% edges are kept ($\alpha = 0.2$). During this process the robustness of the scaling pattern is also evidenced by $D < 0.035$. According to this result, we can predict the scaling pattern under the condition that more clickstreams were collected: the value of γ would be smaller and the fitting is likely to be better.

In the second analysis we randomly shuffle the clickstream networks (in the eight different ways given by Tab. 3) and then observe the robustness of the scaling relationship.

The "randomly shuffled links" is different from the "randomly connected links" in Table 3. The former keeps the long-tail degree distribution of the original network, but the latter leads to a ER random graph model of binomial degree distribution [1]. Similarly, the "randomly shuffled weights" is different from the "uniformly distributed weights", for we permute the order of weights and keep

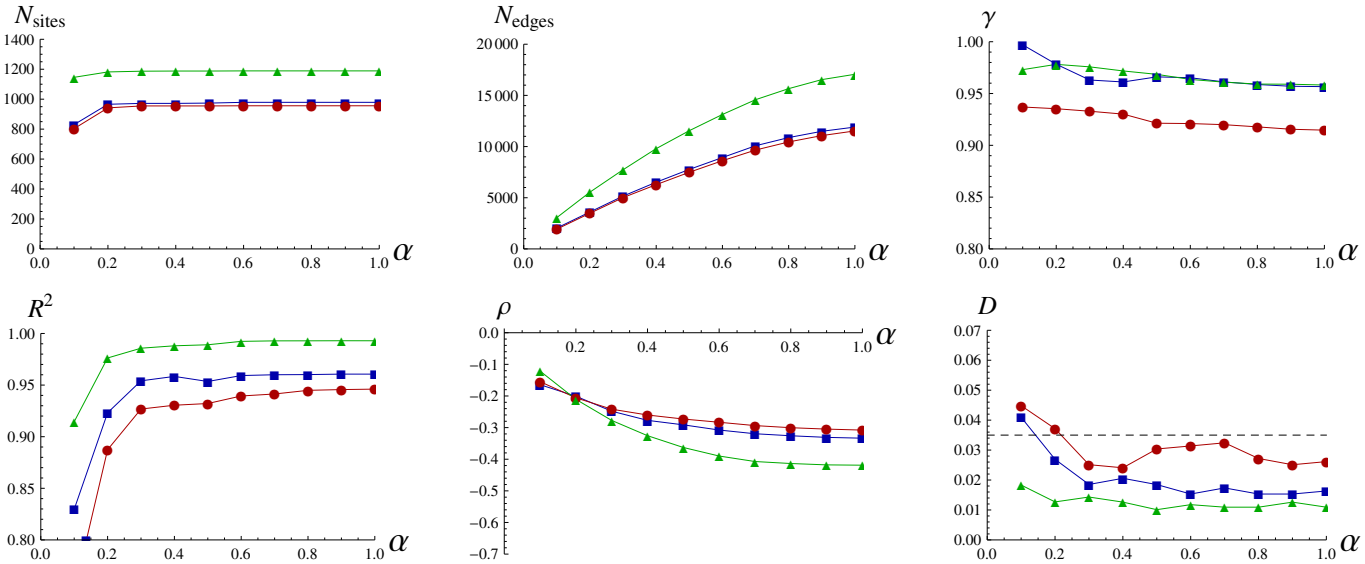

Fig. 6. The change of the number of nodes, the number of links, γ, R^2, correlation, and KS statistic with the increase of α. The data points from three networks are plotted in different colors: blue squares for $w1$, red circles for $w2$, and green triangles for $w3$. The dashed, black line in the last figure shows the critical value of KS statistics (0.035).

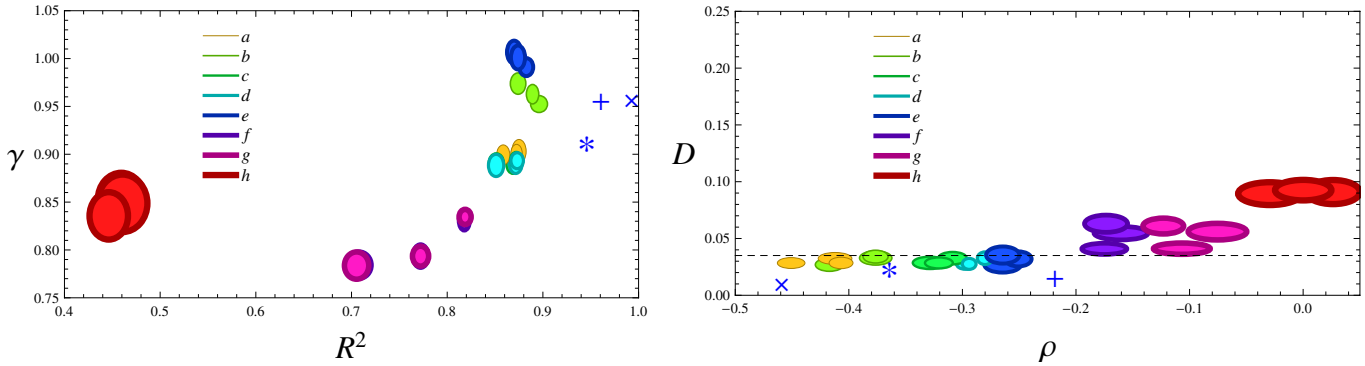

Fig. 7. The mean and standard deviation of statistics of interested of the eight combinations in the reshuffling. The dashed, black line in the left figure shows the critical value of KS statistics as 0.035 given the condition of 1200 sample size and 0.1 level of confidence.

Table 3. The combinations in the reshuffling.

	Ori weights	Ran. shuffled weights	Uni. distributed weights
Ori. links	$w1/w2/w3$	a	b
Ran. shuffled links	c	d	e
Ran. connected links	f	g	h

their long-tail distribution in the former, but create new, uniformly distributed weights in the latter.

For each of the combinations listed in Table 3, we ran 100 times of simulations and recorded the mean and standard deviation of the aforementioned four statistics. After that, we plotted γ vs. R^2 and ρ vs. D (Fig. 7). In Figure 7, the center of the disks indicates the mean values and the radius reflects the standard deviations. We plot the results of different combinations in distinct colors and show the results of the empirical networks by "+" ($w1$), "$*$" ($w2$), and "$\times$" ($w3$). It turned out that across the three studied networks, the original networks always had the smallest D and largest R^2. We find that "randomly connected links" usually leads to $D > 0.035$. Therefore, we can conclude that the discussed scaling pattern is related to the topological structure of the clickstream networks rather than the distribution of weights on clickstreams.

4 Discussion

We study collective browsing behavior from a flow network perspective. We define C_i as a measure of the impact of websites i on other sites through users' collective, continuous surfing activities and found it scaled to website traffic A_i with an exponent smaller than 1. This

pattern unrevealed the decentralized structure of the three clickstream networks under study. Further, we found that this scaling pattern appeared universally across language-based communities with a γ independent of communities size. Finally, we examined the stability of the scaling pattern against the permutations of the clickstream networks. It turned out that the scaling relationship was robust against the selective removal of edges but sensitive to the change of the linking structure.

Our finding has relevant theoretical and practical consequences. Although the "rich-get-richer" paradigm has been widely accepted as a mechanism of hyperlink formations since Barabási and Albert [13], we should not simply assume that this paradigm also suits the dynamics of collective surfing behavior [14,15]. It is already pointed out in reference [16] that the traffic of websites scaled to its number of inbound links with an exponent approaching 0.8. In this work we found the sublinear relationship between the impact and the traffic. Puting these findings together, we can conclude that the survival probability of small sites in the web ecological system is actually higher than what was suggested by their in-degree [13] or the Page Rank values based on hyperlink structure [5,16]. Moreover, we would like to emphasize that it is only by studying empirical clickstream networks can these conclusions be achieved.

Finally, the found scaling relationship provide a quantitative prediction of the impact of a website from its traffic. Online advertising usually measures the impact of websites by their traffic [35,36], but our study offers a more precise calculation of the impact of sites based on their role in the circulation of clickstreams. Therefore, this approach has potential application in the estimation of the value of sites and also the planning of online marketing campaigns.

We acknowledge the support from the National Natural Science Foundation of China under Grant No. 61004107.

References

1. D. Watts, S. Strogatz, Nature **393**, 440 (1998)
2. A. Broder et al., Computer networks **33**, 309 (2000)
3. J. Kleinberg, S. Lawrence, Science **294**, 1849 (2001)
4. J. Kleinberg et al., Nature **406**, 845 (2000)
5. L. Page, S. Brin, R. Motwani, T. Winograd, The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab (1999). `http://ilpubs.stanford.edu:8090/422/`
6. M. Meiss, B. Gonçalves, J. Ramasco, A. Flammini, F. Menczer, in *Proceedings of the 21st ACM conference on Hypertext and hypermedia, ACM, 2010*, pp. 229–234
7. F. Qiu, Z. Liu, J. Cho, in *Proceedings International Workshop on the Web and Databases (WebDB), Citeseer 2005*, pp. 103–108
8. A. Chmiel, K. Kowalska, J. Hołyst, Phys. Rev. E **80**, 066122 (2009)
9. R. White, J. Huang, in *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval, ACM 2010*, pp. 587–594
10. M. Meiss, F. Menczer, S. Fortunato, A. Flammini, A. Vespignani, in *Proceedings of the international conference on Web search and web data mining, ACM, 2008*, pp. 65–76
11. B.A. Huberman, P. Pirolli, J. Pitkow, R. Lukose, Science **280**, 95 (1998)
12. J. Bollen et al., PLoS One **4**, e4803 (2009)
13. A. Barabási, R. Albert, Science **286**, 509 (1999)
14. J. Cho, S. Roy, in *Proceedings of the 13th international conference on World Wide Web, ACM, 2004*, pp. 20–29
15. L. Introna, H. Nissenbaum, Computer **33**, 54 (2000)
16. S. Fortunato, A. Flammini, F. Menczer, A. Vespignani, Proc. Natl. Acad. Sci. **103**, 12684 (2006)
17. J. Brainerd, B. Becker, in *Proceedings of the IEEE Symposium on Information Visualization, 2001* (INFOVIS'01), IEEE Computer Society, p. 153
18. G. Funkhouser, M. McCombs, The Public Opinion Quarterly **35**, 107 (1971)
19. K. Lerman, R. Ghosh, in *Proceedings of 4th International Conference on Weblogs and Social Media (ICWSM), 2010*
20. F. Wu, B.A. Huberman, Proc. Natl. Acad. Sci. **104**, 17599 (2007)
21. C. Cattuto, A. Barrat, A. Baldassarri, G. Schehr, V. Loreto, Proc. Natl. Acad. Sci. **106**, 10511 (2009)
22. F. Wu, D. Wilkinson, B. Huberman, in *Computational Science and Engineering, 2009. CSE'09.* International Conference on IEEE, **4**, 409 (2009)
23. N. Foti, J. Hughes, D. Rockmore, PloS One **6**, e16431 (2011)
24. T. Fruchterman, E. Reingold, Software: Practice and experience **21**, 1129 (1991)
25. J. Zhang, L. Guo, J. Theor. Biol. **264**, 760 (2010)
26. M. Barber, Ecological Modelling **5**, 193 (1978)
27. M. Higashi, Ecological Modelling **32**, 137 (1986)
28. D. Garlaschelli, G. Caldarelli, L. Pietronero, Nature **423**, 165 (2003)
29. D. Warton, I. Wright, D. Falster, M. Westoby, Biol. Rev. **81**, 259 (2006)
30. P. Pirolli, *Information foraging theory: Adaptive interaction with information* (Oxford University Press, New York, 2007), Vol. 2
31. M. Higashi, B. Patten, T. Burns, Ecological modelling **66**, 1 (1993)
32. S. Vitali, J. Glattfelder, S. Battiston, PloS One **6**, e25995 (2011)
33. N. Smirnov, Annal. Math. Stat. **19**, 279 (1948)
34. A. Clauset, C. Shalizi, M. Newman, Soc. Ind. Appl. Math. Rev. **51**, 661 (2009)
35. D. Rosen, E. Purinton, J. Business Res. **57**, 787 (2004)
36. G. Tan, K. Wei, Electron. Commerce Res. Appl. **5**, 261 (2007)

Supplementary Material

Clickstream network statistics

To construct clickstream networks we prepared three lists of seed site. For convenience we simply called these lists "list1", "list2", and "list3", and the constructed networks "$w1$","$w2$", and "$w3$". Table S1 provided the detailed information of the three lists. The traffic distributions of sites in the lists are shown by Figure S1. From Figure S2 to Figure S4, we showed several statistics, including degree, weights, and weighted degree of the clickstream networks.

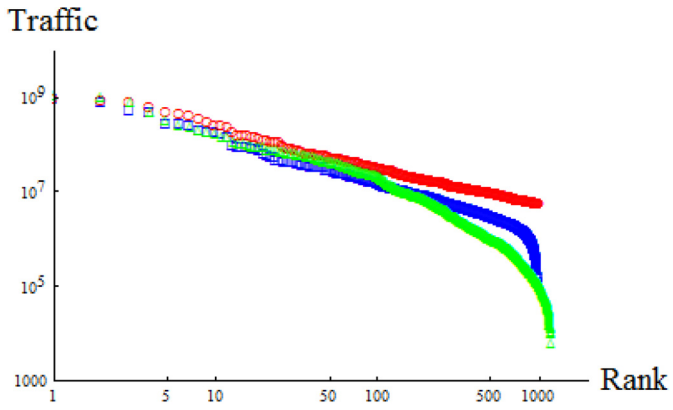

Fig. S1. The traffic distributions of the seed sites. The y-axis is the traffic of sites (measured by the number of unique visitors) and the x-axis denotes the decreasing rank of traffic. In list1 (blue squares) and list2 (red circles), we showed monthly traffic of sites, as the top 1000 site provided by Google is monthly based; but in list3 (green triangles), we used daily traffic for Alexa only offers daily traffic of websites. Both of the x- and y-axes are shown in the base-e log scale.

Table S1. The three lists of seed sites.

List	Collected time	Source	Collected criterion	N_{sites}	Traffic range
list1	Oct., 2010	Google	Top 1000 worldwide	1001	$1.20 \times 10^5 \sim 8.98 \times 10^8$
list2	Jul., 2011	Google	Top 1000 worldwide	1001	$5.50 \times 10^6 \sim 9.00 \times 10^8$
list3	Apr., 2012	Alexa	Top 25 in 124 countries(regions)	1198	$6.13 \times 10^3 \sim 1.11 \times 10^9$

Note: the traffic of sites is measured in number of unique visitors.

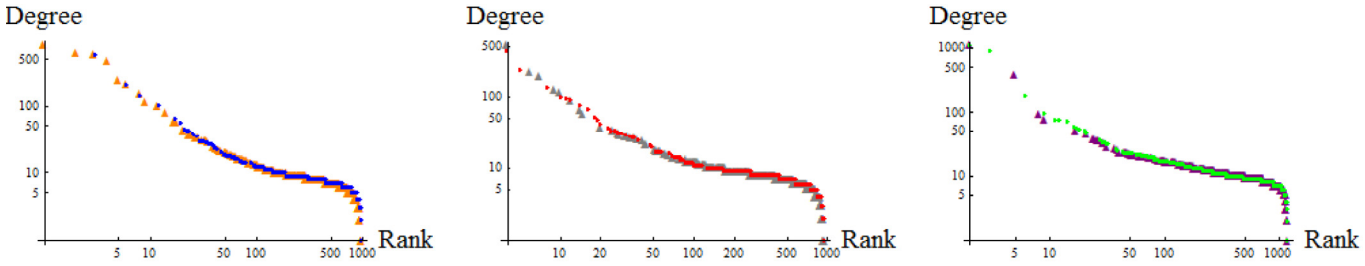

Fig. S2. The degree distributions of the clickstream networks. The y-axis is the degree of nodes (triangle correspond to in-degree and circles correspond to out-degree) and the x-axis denotes the decreasing rank of degree. Both axes are shown in the base-e log scale. The out-degree of sites are plotted in circles (blue for $w1$, red for $w2$, and green for $w3$) and in-degree shown in triangles (orange for $w1$, gray for $w2$, and purple for $w3$).

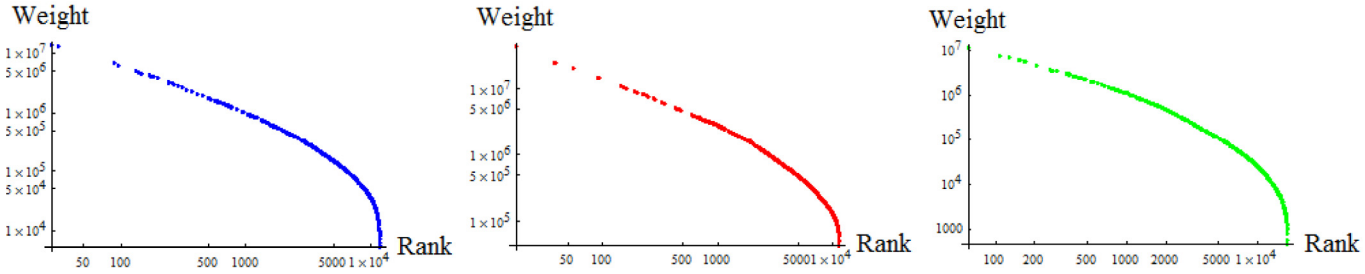

Fig. S3. The distribution of weights in the clickstream networks. The y-axis is the weights of edges and the x-axis denotes the decreasing rank of weights. Both axes are shown in the base-e log scale. The weights in $w1$, $w2$, and $w3$ are shown in blue, red, and green circles, respectively.

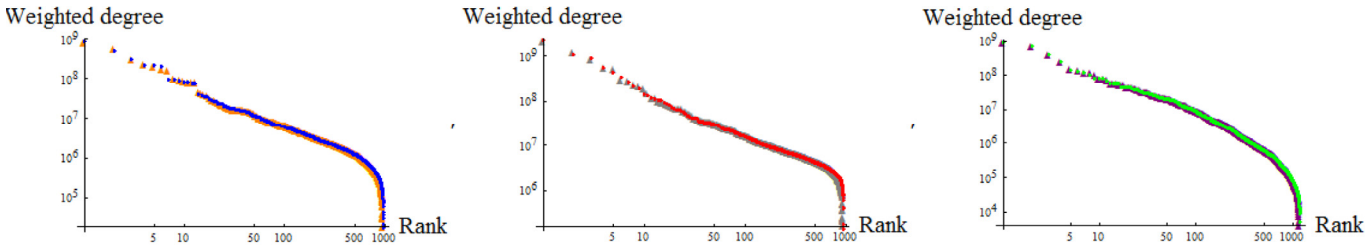

Fig. S4. The weighted degree distributions of the clickstream networks. The y-axis is the weighted degree of nodes (triangles correspond to in-degree and circles correspond to out-degree) and the x-axis denotes the decreasing rank of weighted degree. Both axes are shown in the base-e log scale. The out-degree of sites are plotted in circles (blue for $w1$, red for $w2$, and green for $w3$) and in-degree shown in triangles (orange for $w1$, gray for $w2$, and purple for $w3$).

Language community analysis

We placed, here, two tables showing the statistics of language communities in $w1$ and $w3$.

Table S2. The quantities of interest across language communities in $w1$.

Community	N_{sites}	N_{edges}	Daily clickstreams	γ	R^2 of γ
English	518	5893	3.74×10^9	0.97	0.92
Chinese	208	1987	7.72×10^8	0.93	0.85
Japanese	99	924	2.39×10^8	0.94	0.93
German	30	135	1.20×10^7	0.87	0.88
Russian	28	163	5.46×10^7	1.05	0.87
Korean	27	225	1.21×10^7	0.96	0.89
French	27	72	3.39×10^6	0.94	0.82
Italian	10	25	3.61×10^6	1.19	0.85
Portuguese	9	36	8.54×10^6	0.89	0.92
Vietnamese	7	28	1.80×10^6	0.94	0.93
Polish	5	16	3.55×10^6	0.82	0.88
Thai	3	6	3.12×10^5	0.50	0.99

Note: the daily clickstreams is obtained by summing up the number of users over all edges within a community.

Table S3. The quantities of interest across language communities in $w3$.

Community	N_{sites}	N_{edges}	Daily clickstreams	γ	R^2 of γ
English	339	4157	3.62×10^9	0.99	0.97
Arabic	116	700	2.71×10^7	0.94	0.96
Spanish	112	535	4.27×10^7	0.99	0.97
Russian	66	393	9.11×10^7	1.00	0.97
French	63	254	3.55×10^7	1.02	0.98
Chinese	50	349	5.90×10^8	1.29	0.93
Croatian	33	188	3.79×10^6	1.10	0.93
German	30	192	5.25×10^7	1.00	0.97
Czech	26	177	1.90×10^7	1.01	0.83
Persian	21	78	4.81×10^6	0.98	0.91
Romanian	19	86	4.83×10^6	1.13	0.91
Polish	18	141	2.85×10^7	0.95	0.93
Japanese	18	137	1.65×10^8	1.04	0.94
Hungarian	17	103	7.36×10^6	0.94	0.94
Greek	17	84	4.45×10^6	1.13	0.96
Turkish	16	77	1.88×10^7	0.97	0.98
Dutch	16	83	1.09×10^7	0.98	0.94
Portuguese	14	64	2.50×10^7	0.95	0.97
Macedonian	14	67	1.72×10^5	0.83	0.68
Lithuanian	14	61	6.18×10^5	1.09	0.80
Vietnamese	13	41	2.68×10^6	0.91	0.75
Italian	13	63	2.78×10^7	0.95	0.95
Finnish	13	83	5.35×10^6	0.91	0.94
Norwegian	12	81	5.75×10^6	0.87	0.86
Estonian	11	61	9.45×10^5	0.96	0.85
Icelandic	10	59	5.18×10^5	1.03	0.96
Latvian	9	52	1.30×10^6	0.93	0.93
Hebrew	9	46	2.52×10^6	1.00	0.94
Danish	9	49	3.37×10^6	0.91	0.97
Bulgarian	9	36	5.86×10^5	0.78	0.82
Albanian	8	42	2.61×10^5	0.86	0.88
Thai	7	23	9.42×10^5	0.82	0.34
Georgian	7	13	9.93×10^4	0.96	1.00
Azerbaijani	6	14	1.30×10^5	0.95	1.00
Slovenian	5	11	1.07×10^5	1.08	0.71
Korean	5	16	3.88×10^6	0.92	0.82
Slovak	3	4	9.83×10^4	0.50	0.89

Note: the daily clickstreams is obtained by summing up the number of users over all edges within a community.

The example results of the backbone network analysis

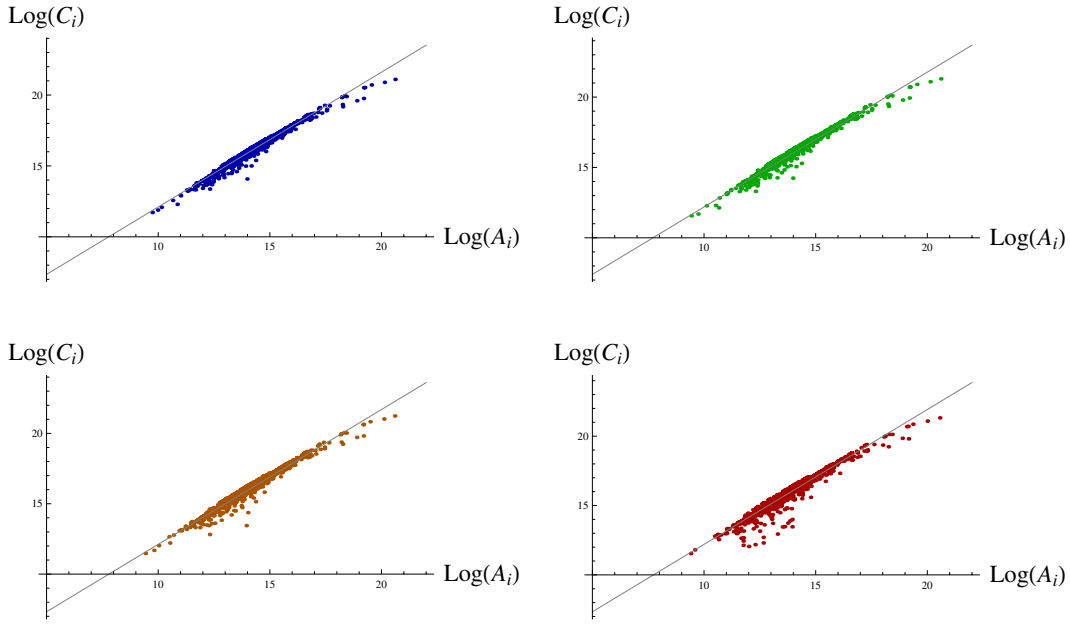

Fig. S5. The example steps in the skeleton analysis applied on $w1$. The blue, green, yellow, and red points correspond to nodes within the backbone networks when $\alpha = 0.8$, $\alpha = 0.6$, $\alpha = 0.4$, and $\alpha = 0.2$, respectively. Both of the x- and y-axes in the four figures are shown in the base-e log scale.

The example results of the reshuffling analysis

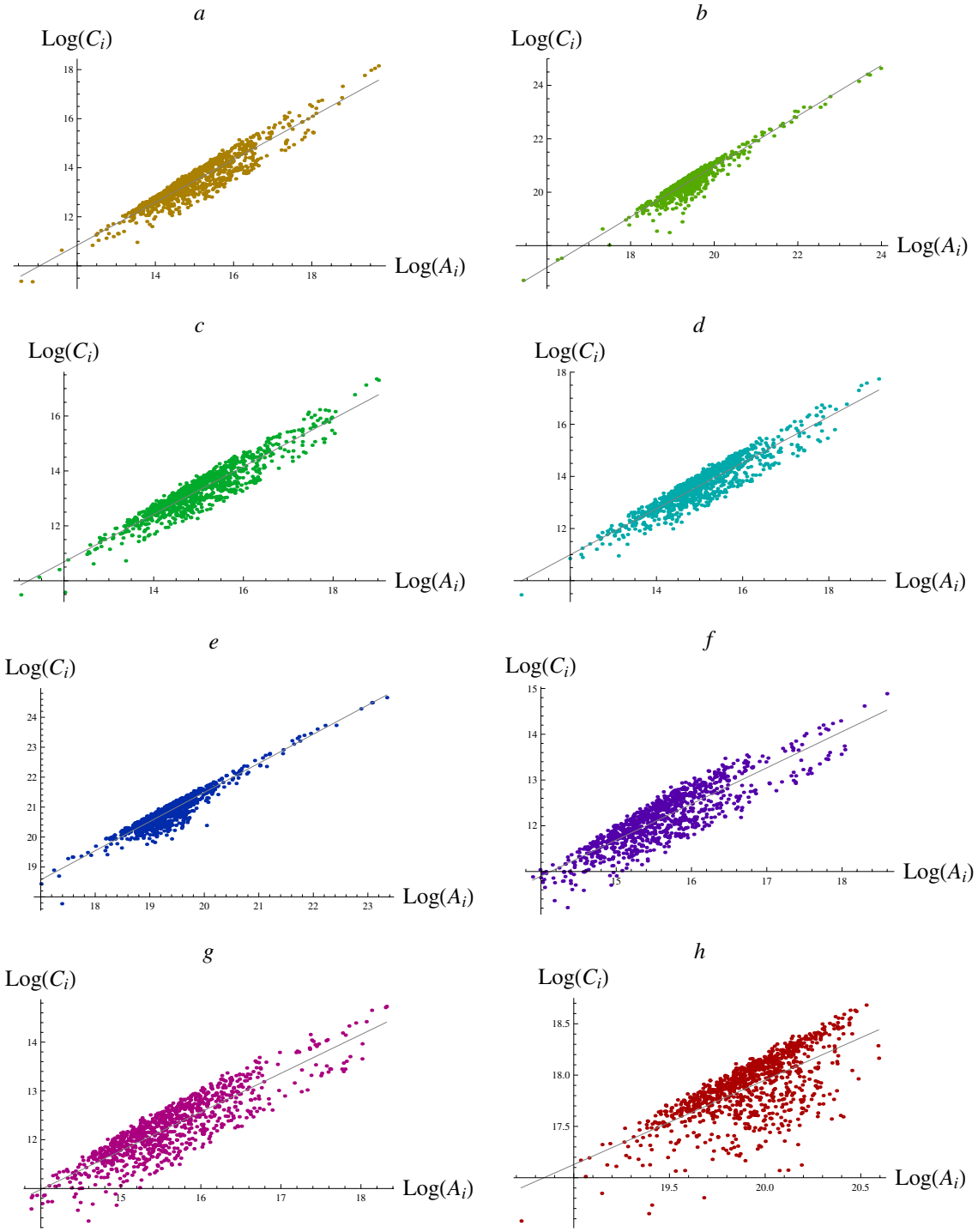

Fig. S6. The example results of the eight combinations in the shuffling analysis applied on $w1$. The data points of different combinations are shown in distinct colors. Both of the x- and y-axes in the eight figures are shown in the base-e log scale.