SCIENCE CHINA
Information Sciences

• **Supplementary File** •

Common patterns of online collective attention flow

Yong LI[1,2], Xiaofeng MENG[1*], Qiang ZHANG[2], Jiang ZHANG[3] & Changqing WANG[4]

[1]*School of Information, Renmin University of China, Beijing* 100872*, China;*
[2]*College of Computer Science and Engineering, Northwest Normal University, Lanzhou,* 730070*, China;*
[3]*School of Systems Science, Beijing Normal University, Beijing,* 100875*, China;*
[4]*DNSLAB, China Internet Network Information Center, Beijing,* 100190*, China*

Appendix A Background and related works

With the arrival of the era of big data, availability of those from human online surfing records and communication records, accompanied by a wide range of high-throughput measurement tools and technologies, data analytics has become the hottest industry. It not only impacts the subject areas from computer science to sociology, but also gives the emergence of new scientific disciplines such as social computing [1] and computational social science [2]. Although such online behavioral data is called "small data" in the era of big data [3], it can help explain many complex socio-economic phenomena.

Traditional studies focused on the information flow on the Web, however, we would look at this problem from an "inverse" perspective, which is to study from the attention flow. We define the attention flow as an ordered sequence of webpages viewed by a user and study the weighted collective attention flow network, which is a collaborative product of massive number of online users.

As pointed out by Herbert Simon[1)]: "a wealth of information creates a poverty of attention", attention will play a more important roles in the future because of the overload of information and the scarcity of attention [4]. In a recent study, we found that the online collective attention flow can be used to quantify the influence of websites. It is an effective theoretical tool to estimate and rank websites [5]. However, we still don't know the common patterns of these behaviors. It is thus of interest to probe into underlying mechanisms how collective attention flow allocates among the over-abundant information source and impacts evolutionary of the Web [6].

In ecosystem, energy flow is one of the most important subject. As shown in Figure S1, in 1932, Kleiber found that the "metabolism"(M_b) and "body mass"(B) of a living organism follows a universal power law distribution with the exponent being about 3/4, i.e., $M_b \propto B^{3/4}$ [7]. For example, an elephant weights 10000 times more than a mouse, but the energy required for an elephant is only 1000 times than that of a mouse. Recently, researchers found that all flow systems follow a similar allometric scaling law just like metabolic systems [8]. The allometric scaling laws of energy flow networks may explain the rule of collective attention flow network.

Different from previous studies which attempted to figure out how users surf the Internet and how information flow transmitted, we want to understand how websites consume collective attention flow in this paper. Inspired by metabolic theory, we view the Web as virtual living tissues that grow at the cost of online collective users' attention flow. We quantify the metabolic rate of websites, and obtain the Web version of Kleiber's law [7], which describes how the collective attention as energy is fed into sites and dissipated out of sites.

As one of the main methods in network science [9], complex network is a useful tool to study interactions and relationships between components of a complex system. Recently, by incorporating statistical mechanics and graph theories, scholars have uncovered a series of universal patterns in various weighted networks such as the metabolism networks [10], world trade web [11], citation networks [12], technological networks [13, 14], and the complex dynamical networks [15]. The new common patterns include allometric scaling laws in ecology [10], the long tailed distribution of total weights of each node [16], the linear relationship between the power law exponents of in-degree and out-degree [17], and power law strength-degree correlation from resource-allocation [18]. Several models have been proposed to explain these patterns such as generative model [12], traffic-driven evolution model [13], mutual selection model [14], time-varying complex dynamical

* Corresponding author (email: xfmeng@ruc.edu.cn)

1) Herbert Simon, (1916-2001), the Turing Award 1975 and the Nobel Prize 1978 winner.

network model [19], etc. However, these patterns and models rarely consider long-ranging, complex interactions between the Web and users.

In recent years, scholars have paid a special attention to online collective behaviors based on network big data. Zhao et al. [20] investigate the scaling behaviors associated with human-interest dynamics based on collective surfing data from Douban[2)], Taobao[3)], and Mobile-Phone Reading[4)]. They quantified human-interest based on basic variables such as interest length, return time, and interest transition and found that there are three basic ingredients underlying human-interest: preferential return, inertial effect, and exploration. In addition, Wu et al. [21] obtained data from Baidu Tieba[5)], the biggest Chinese searching engine and forum site. They studied collective browsing behaviors from the clickstream network's perspective and discovered the allometic scaling law between page views and the number of unique visitors in a given time period.

However, in contrast to human-interest, attention flow is a more general concept of human online behaviors. Furthermore, these studies only focus on a single site, especially in E-commercial site. Their datasets are sites-centric data, which do not contain individual successive surfing records among the entire World Wide Web. Thus, these studies is only the micro-level dynamics of human online behaviors and cannot be used to look into the common patterns of online collective attention dynamics.

Appendix B Data and Methods

Data. We obtain data from China Internet Network Information Center(CNNIC)[6)], an operation, administration and service organization of national network fundamental resources of China. The data have been collected from more than 30000 online volunteers' surfing records for more than five years. Every volunteer install a client data acquisition program on his or her personal computer. The detailed records of the user's online behavior include information such as the timestamp of each focus screen window, the window process name, the URL address, and user's demographic attribute, etc. CNNIC release a report of Chinese Internet development analysis every half a year mainly based on this dataset , which has been widely accepted and cited in the academic and industrial community.

We randomly sample 1000 volunteers' surfing data within a month and obtain about 120 million(1.2×10^8) records. We analyze the sampled data and confirm the common patterns of online collective behaviors. Based on the results of sampled data analysis, we have conducted the data analysis programs to the server of CNNIC to validate the results for all users and for all time. From the view of statistical theory, 30000 online volunteers cannot represent about more than 600 million Chinese Internet netizens, but the surfing records of more than 30000 volunteers over five years can accurately describe the common patterns of online collective behaviors. Like the study of the patterns of the swarm (ants and bees), we only need to focus on the patterns of the collective behavior from a systems science's perspective.

Attention flow network. To obtain a weighted collective attention flow network from our dataset, we first transform the users surfing records into "click lists" involving the timestamp and domain names for each HTTP request. Consecutive clicks of webpages in the same site in the same session by a user are considered as a single attention flow online. The resulting click lists records are a collection of collective users' attention flow between sites that shows the transformation of collective attention flow from sites to other sites. We apply a common practice of a session for users' dwelling time in a site, by setting the threshold of a session in thirty minutes [22]. We make the assumption that the user time between two sessions is not spent online, while the time during a session is spent online.

As shown in Figure S2, we establish a site-level attention flow network. The network is a weighted directed graph with nodes being the sites and the edges being the collective users' attention flow between sites. Node size is proportional to the collective users' dwelling time, and edge weight is proportional to the value of attention flow traffic intensity between two sites.

Formally, we write the collective attention flow network as a weighted directed graph G, denoted by

$$G = (V, E, T, W).$$

V indecates nodes set of size $n+2$, $E \subseteq V \times V$ denotes a set of edges representing attention flow traffic intensity between sites, T is the set of collective users' dwelling time of each node. W represents the set of edges weight and are related to E. The value of W is the magnitude of attention flow traffic between sites, indicating the weight of the edges(the weight of a non-existing edge is defined to be zero).

We add two special artificial vertices, "source" and "sink" (node 0 and node $n+1$), as shown in Figure S2, representing the source and the sink of the attention flow respectively. A user starts his or her online trip from "source", while he or she stops surfing when a session ends, resulting in that the exported attention flow dissipates to "sink" node.

Basic variables. From a given graph G, we define a weighted matrix M as a flow matrix which represents an attention flow network:

$$M_{(n+2)\times(n+2)} = \{W_{ij}\}_{(n+2)\times(n+2)},$$

where W_{ij} is the weight of edge from node i to node j in graph G, $\forall i, j \in [0, n+1]$. Let m_{ij} represents the element of flow matrix M. Note that the diagonal element $m_{ii} = 0$.

2) http://www.douban.com/
3) http://www.taobao.com
4) MPR. a widely used electronic reading tool
5) http://tieba.baidu.com
6) http://www.cnnic.cn/

As shown in Table 1 of letter file, we define six basic variables related to weight of each site to infer the common patterns of co-evolution of the Web and online collective attention flow. According to the flow matrix M, we can estimate the total flow size of a given site i. This value is also commonly known as the node strength in complex weighted network literature [10], mathematically,

$$A_i = \sum_{j=1}^{n+1} m_{ij} = \sum_{j=0}^{n} m_{ji}, \forall i \in [1, n],$$

where $\sum_{j=1}^{n+1} m_{ij}$ is the total outflow and $\sum_{j=0}^{n} m_{ji}$ is the total inflow of site i because "source" node (node 0) only has outflow and "sink" node (node $n+1$) only has inflow. Notice that the attention flow network is balanced because a user entering a site will for sure leave that site after some time in a session, meaning that the total inflow equals the total outflow for each site. We only need to estimate the inflow or outflow of each site.

We define two vectors T_i and P_i, to represent the weight of a node i, where $T_i = \sum_{j=1}^{k} t_j$. Suppose there are k users browsing site i during the observing duration, every user's dwelling time is t_j, thus T_i represents the overall dwelling time of all users on a site i. We define the temporal of vertices "source" and "sink" being zero. The vector P_i represents the overall page views of a site i on an observing duration. The vector D_i is the node degree(the total number of inward edges and outward edges).

In addition, we define two variables I_i and H_i, where I_i is the attention flow intensity of site i from "source"(node 0): $W_{0,i}$, and H_i is the dissipated attention flow to "sink"(node $n+1$). For example, in example attention flow network in Figure S2, I_1=50, H_6=10 and D_2=4.

DGBD distributions. We use the discrete version of a generalized beta distribution(DGBD) [23] to fit the distributions of six basic variables in Table 1 of letter file. DGBD is a rank-ordered distribution, which has two parameters and do not need to be given the empirical density function of distribution. The performance of DGBD fitting is so outstanding that it is suitable for many diverse disciplines, such as social sciences, natural sciences and arts. The functional form of DGBD is defined as:

$$p(r_i) = \frac{k(N+1-r_i)^b}{r_i^a}, (a, b) \geqslant 0, \tag{B1}$$

where $p(r_i)$ is the value of the variable distribution, and r_i is the ranking value of site i and are sorted in a decreasing manner. N is the total number of nodes in the attention flow network; (a, b) are two fitting exponents and k is simply a normalization constant. The exponent a is interrelated with the power law phenomena of the concerned variable, while b is related to the discrete range for disordered fluctuations.

If we set $b = 0$ in Eq. (B1), then $p(r_i) = k/r_i^a$, it is the distribution function of Zipf's law [24] which is the frequency of words used in a specific language. In Zipf's law, the logarithm of word frequencies had a linear relationship with respective to the logarithm of their ranks after the word frequencies being sorted in a decreasing order where the slopes is -1, which implied a power law phenomenon. However, this power law behavior holds only within a partial breaking negative slope straight line. In order to fix this problem, Martnez-Mekler, et al. [23] adopted a new exponent b in the DGBD fitting. Thus, by tuning the value of the extra exponent b, DGBD can be used to fit not only the data with power laws distribution but also the data with an obvious deviation from power law distribution.

We fit six basic variables of attention flow network by DGBD curves with the OLS(ordinary least square) regression. Figure S3 shows the distribution of each variables, where the green solid line is DGBD fitting for T_i with parameters (k, a, b)=(13.12, 1.8380, 0.9083); the blue dashed line is DGBD fitting for P_i with parameters (k, a, b)=(14.56, 1.6129, 0.1837), and the red solid line is DGBD fitting for A_i with parameters (k, a, b)=(13.73, 1.4145, 0.198). R^2 is the explained variance of the regression. The R^2 values of these three fitted variables are all above 0.97, which indicates that these fittings are very good for the whole range of rank-ordered distribution values.

In Figure S3, the green dashed line is the nodes degree distribution (D_i) with parameters (k, a, b)=(10.78, 1.0068, -0.0316); the blue solid line is I_i with parameters (k, a, b)=(7.19, 0.5579, -0.2313); and the red dashed line is DGBD fitting for H_i with parameters (k, a, b)=(8.23, 0.6549, -0.2441). It is observed that the distribution of inflow from "source"(I_i) is similar to the distribution of dissipated outflow to the node "sink" (H_i) in an attention flow network. Although these three fittings are very good for the whole distribution of data, however, the exponent b of three curves are smaller than zero, which are not consistent with the definition of DGBD.

Clearly, in Figure S3, there are two crooked points in every fitted curve, and thus dividing the curve into three parts, where the left and right parts of a curve are much steeper than the middle one. Note that each part yields a logarithmic decreasing phenomenon. Given the DGDB distribution of six variables we know that there is a large fluctuation from the left to the right parts of the curves. The DGBD distribution demonstrate how a process decays for a given property. The slopes of the curves indicate the heterogeneities of the property correspondingly. We can conclude that the nodes in the left and right parts of the distribution curves are more heterogeneous than the ones in the middle. Based on the similarity of the distribution curves in Figure S3, we make an educated guess that there may exist a connection between the basic variables. The analysis reveals several surprising findings in online collective attention flow network, such as allometric scaling laws, dissipation laws, gravity laws, and Heap's law.

Appendix C Meaning of allometric scaling laws exponent α

In previous studies of allometric scaling laws on ecological flow network and clickstream network [21, 29], the exponent α can be explained as the level at which large nodes dominate the circulation on the entire network. This explanation is also consistent with attention flow network. For example, as shown in $A_i \sim D_i^{\alpha}$, we have three sites with the different

node degrees D_i being $\{1, 2, 3\}$, with the different values of α. They have different attention flow intensity in the network; the higher α, the higher attention flow intensity will be. Assume α_1 is 0.5, then $A_i \sim D_i^{0.5} = \{1, 1.4, 1.7\}$, the third site dominates $1.7/(1 + 1.4 + 1.7) = 41.5\%$ of attention flow. When α_2 is 1, then $A_i \sim D_i^1 = \{1, 2, 3\}$, the third site dominates $3/(1 + 2 + 3) = 50\%$ of attention flow intensity. If α_3 is 2, we would have $A_i \sim D_i^2 = \{1, 4, 9\}$, the third site dominates $9/(1 + 4 + 9) = 64.3\%$ of attention flow intensity. The larger α is, the more centralized the attention flow intensity will be.

To conclude, α greater than one implies that the attention flow intensity is centralized in a small group of hub nodes in attention flow network. These hub sites control the attention flow of the whole network and a small node degree can provide much attention flow intensity in the whole network. α less than one means the attention flow on the network is evenly distributed, and thus the attention flow network is more decentralized. This conclusion is also applied to other basic variables, such as P_i, T_i, etc.

Appendix D Meaning of dissipation laws exponent β

In the perspective of Kleiber's law [7], if we treat the Web systems as an integrated organism, the dissipation H_i would be its metabolism and T_i would be its body mass. By using the dissipation exponents of the sites, we can immediately find some remarkable patterns. We have the dissipation law $H_i \sim kT_i^{\beta}$, where β is 0.81 which is less than one. By differentiating of this function we find that a small increase in H_i per unit increase in T_i:

$$f_1 : \ \frac{dH_i}{dT_i} \propto k\frac{H_i}{T_i} \simeq \frac{T_i^{\beta}}{T_i} = T_i^{\beta-1} = T_i^{-0.19}. \tag{D1}$$

As shown in the solid curve in Figure S9, Eq. (D1) tells us the follows. If we assume that the sites sustain by absorbing the energy of any online collective user life time (T_i), the dissipation (H_i) will decrease as T_i increases. This is called the metabolic rate of the sites, which reflects both the ability of the Web transport system to deliver metabolites to the organisms and the ability that organisms use them. This means that larger sites can be more efficient than smaller sites when absorbing the attention flow as energy. Therefore, this phenomenon implies an "effectiveness of large-scale" pattern.

In addition, if we regard the Web systems as a battery to store the energy, which is collective attention flow, then the energy that per unit of dissipation needs safisfies

$$f_2 : \ \frac{dT_i}{dH_i} \propto T_i^{1-\beta} = T_i^{0.19}. \tag{D2}$$

As shown in the dashed curve in Figure S9, in ecology, Eq. (D2) shows how much time a site needs to complete the metabolism. We know that time is reciprocal of frequency [8], thus Eq. (D2) also shows that the site tends to speed up the internal energy flow transformation in evolution. Therefore, we can deduce that various frequencies of the Web systems is a scaling relationship ($T_i^{0.19}$) with respect to its attention flow, such as the frequencies of updating speed and capability for innovation. In other words, large sites have low rates for updating activity, such as innovation and change, while small sites have fast updating activity.

Appendix E Relationship of exponents between Zipf's law and Heaps' law.

In appendix B, we know that the exponent a of DGBD distribution is same as the exponent of Zipf's law. Zipf's law and Heaps' law are well known coexistence in the context of complex systems. The previous studies indicated that the Heaps' law can be considered as a derivative phenomenon of the Zipf's law, however, the Zipf's law cannot be derived from the Heaps' law [26]. In this article, the DGBD exponent a of total page views of a site(P_i) is 1.4145, which means that the Heaps' exponent $\theta \approx 1/a$. However, we cannot conclude that the Heaps' law is completely dependent on the Zipf's law and what mechanisms resulting in this phenomenon. One ingredient causing such a phenomenon may be the memory and bursty nature of human online behaviors.

References

1 Meng X, Li Y, Zhu J. Social computing in the era of big data:opportunities and challenges. Journal of Computer Research and Development, 2013, 50: 2483-2491
2 Lazer D, Pentland A, Adamic L, et al. Computational social science. Science, 2009, 323: 721-723
3 Estrin D. Small data, where n=me. Communications of the ACM, 2014, 57: 32-34
4 Shi P, Huang X, Wang J, et al. A geometric representation of collective attention flows. PLoS ONE, 2015, 10: e0136243
5 Li Y, Zhang J, Meng X, et al. Quantifying the influence of websites based on online collective attention flow. Journal of Computer Science and Technology, 2015, 30: 1175-1187
6 Wu F, Huberman B A. Novelty and collective attention. Proc Nati Acad Sci USA, 2007, 104: 17599-17601
7 Mackenzie D. New clues to why size equals destiny. Science, 1999, 284: 1607-1609
8 Zhang J. Energy flows in complex ecological systems: a review. Journal of Systems Science and Complexity, 2009, 22: 345-359
9 Barabasi A L. Network science: luck or reason. Nature, 2012, 489: 507-508
10 Zhang J, Feng Y. Common patterns of energy flow and biomass distribution on weighted food webs. Physica A: Statistical Mechanics and its Applications, 2014, 405: 278-288

11 Shi P, Zhang J, Yang B, et al. Hierarchicality of trade flow networks reveals complexity of products. PLoS ONE, 2014, 9: e98247
12 Wang D, Song C, Barabsi1 A L. Quantifying long-term scientific impact. Science, 2013, 342: 127-132
13 Wang W, Wang B, Hu B, et al. General dynamics of topology and traffic on weighted technological networks. Physical Review Letters, 2005, 94: 188702
14 Wang W, Hu B, Zhou T, et al. Mutual selection model for weighted networks. Physical Review E, 2005, 72: 046140
15 Lv J, Wang H, He K. Complex dynamical networks and their applications in software engineering. Journal of Computer Research and Development, 2008, 45: 2052-2059
16 Barabasi A L, Albert R. Emergence of scaling in random networks. Science, 1999, 286: 509-512
17 Liu J, Dang Y, Wang Z, et al. Relationship between the in-degree and out-degree of WWW. Physica A: Statistical Mechanics and its Applications, 2006, 371: 861C869
18 Ou Q, Jin Y, Zhou T, et al. Power-law strength-degree correlation from resource-allocation dynamics on weighted networks. Physical Review E, 2007, 75: 021102
19 Lv J, Chen G. A time-varying complex dynamical network model and its controlled synchronization criteria. IEEE Trans. Automatic Control, 2005, 50: 841-846
20 Zhao Z, Yang Z, Zhang Z, et al. Emergence of scaling in human-interest dynamics. Scientific Reports, 2013, 3: 1-7
21 Wu L, Zhang J, Zhao M. The metabolism and growth of web forums. PLoS ONE, 2014, 9: e102646
22 Kumar R, Tomkins A. A characterization of online browsing behavior. In: Proceedings of the 19th ACM Conference on World Wide Web(WWW2010), Raleigh, 2010. 561-570
23 Martnez-Mekler G, Martnez R A, del Rio M B, et al. Universality of rank-ordering distributions in the arts and sciences. PloS ONE, 2009, 4: e4791
24 Newman M E J. Power laws, pareto distributions and Zipf's law. Contemporary Physics, 2005, 46: 323-351
25 Zhang J, Guo L. Scaling behaviors of weighted food webs as energy transportation networks. Journal of Theoretical Biology, 2010, 264: 760-770
26 Lv L, Zhang Z, Zhou T. Deviation of Zipfs and Heaps Laws in human languages with limited dictionary sizes. Scientific Reports, 2013, 3: 1-7
27 Watts D J, Strogatz S H. Collective dynamics of 'small-world' networks. Nature, 1998, 393: 440-442
28 Huberman B A, Pirolli P L T,Pitkow P J,et al.Strong regularities in world wide web surfing. Science, 1998, 280: 95-97
29 Zhang J, Guo L. Scaling behaviors of weighted food webs as energy transportation networks. Journal of Theoretical Biology, 2010, 264: 760-770
30 Krings G, Calabrese F, Ratti C, et al. Urban gravity: a model for inter-city telecommunication flows. Journal of Statistical Mechanics: Theory and Experiment, 2009, 7:1-8
31 Zhang H. Discovering power laws in computer programs. Information Processing and Management. 2009, 45: 477C483

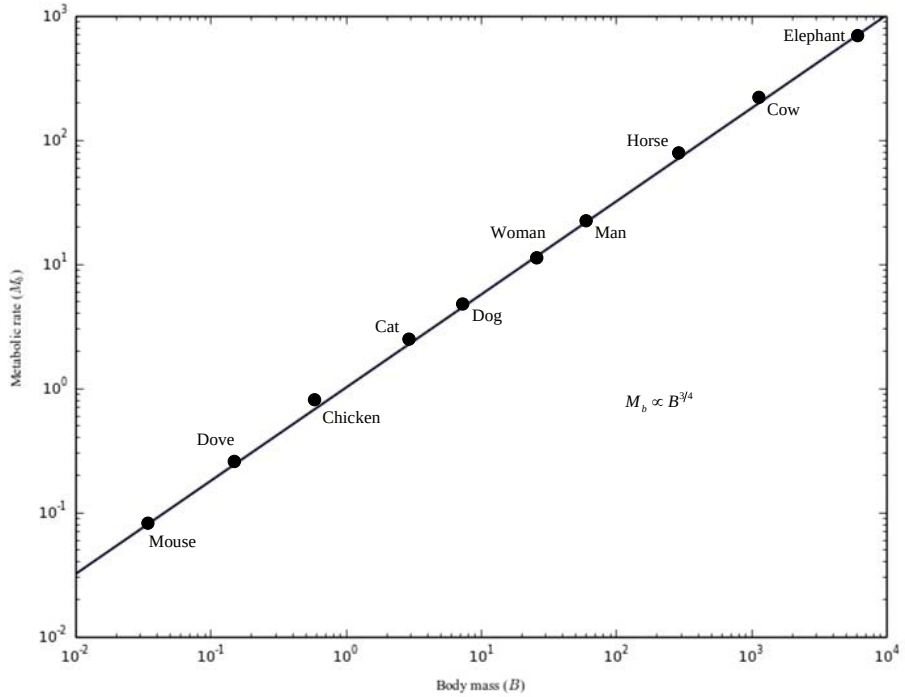

Figure S1 Kleiber's law.

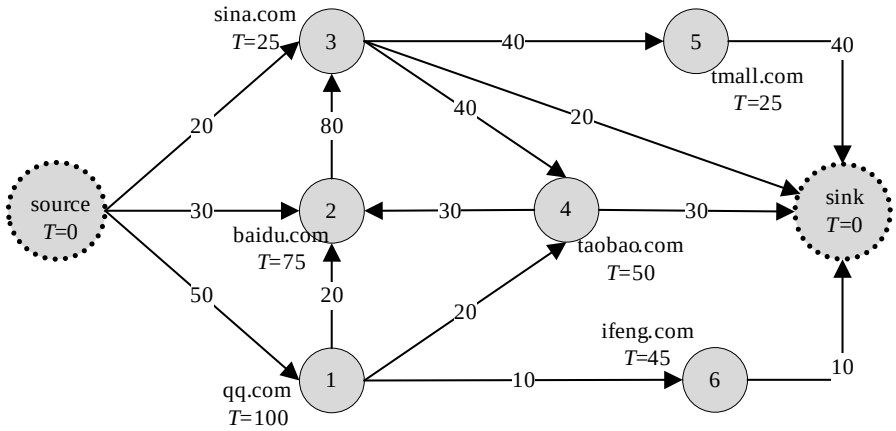

Figure S2 Example of online collective attention flow network.

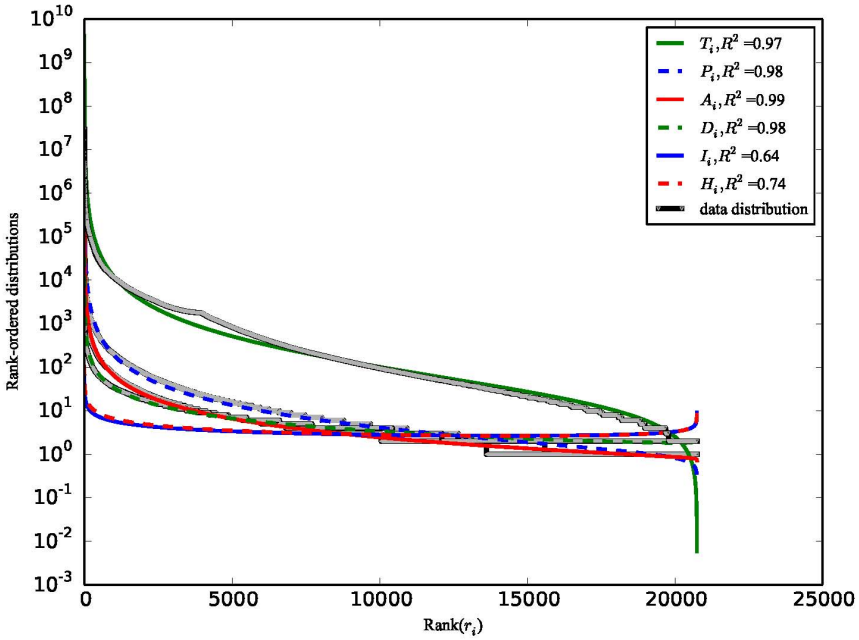

Figure S3 Rank-ordered distributions of basic variables in attention flow network.

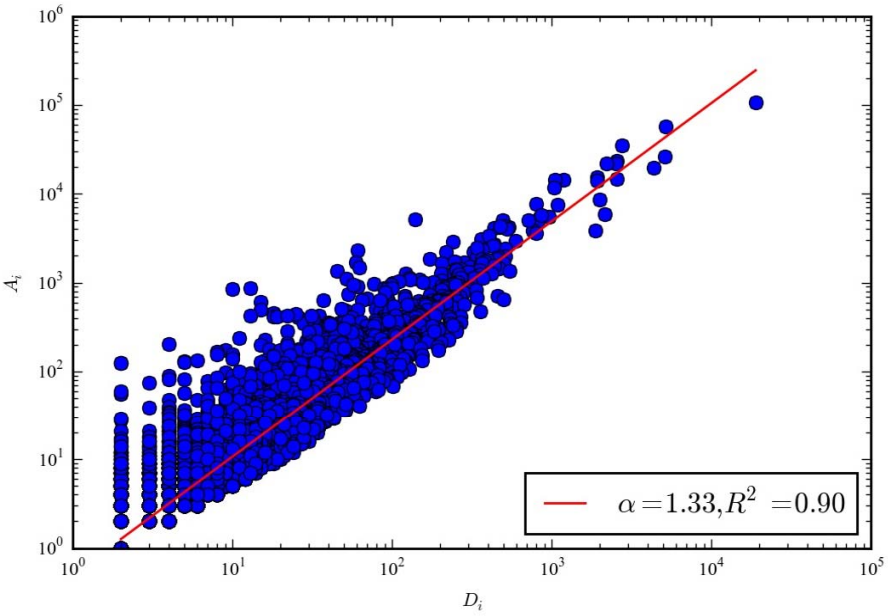

Figure S4 Allometric scaling relationship of $A_i \sim D_i^{\alpha}$ in attention flow network.

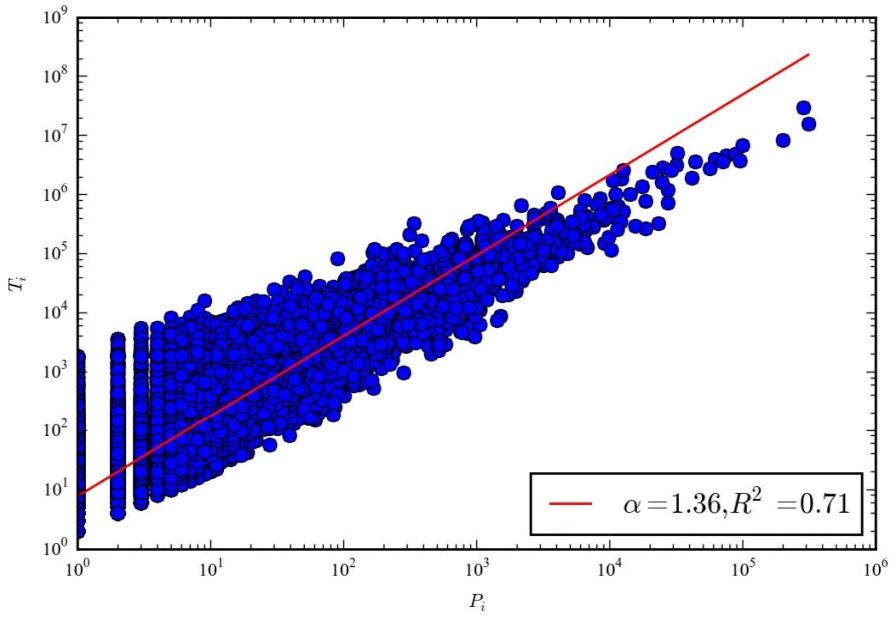

Figure S5 Allometric scaling relationship of $T_i \sim P_i^{1.36}$ in attention flow network.

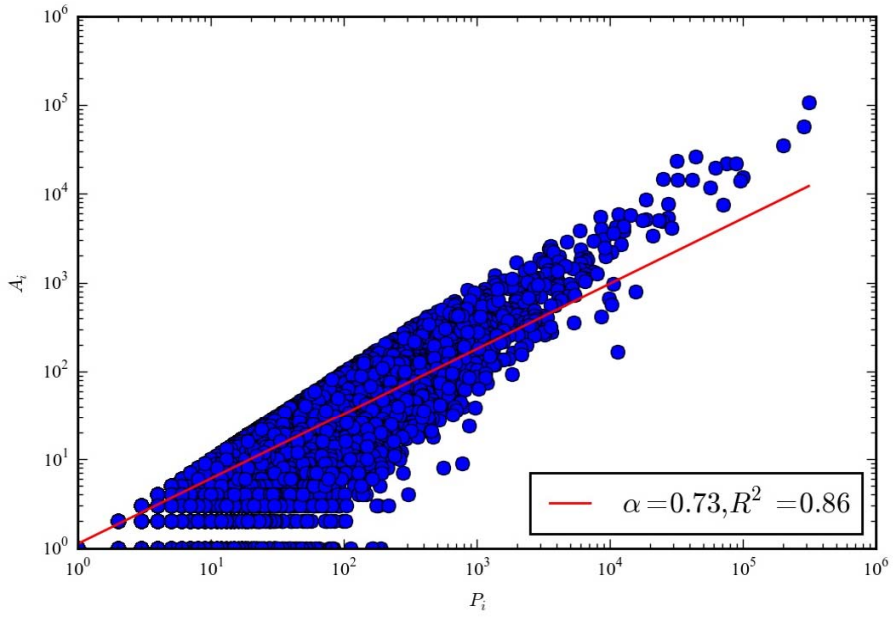

Figure S6 Allometric scaling relationship of $A_i \sim P_i^{0.73}$ in attention flow network.

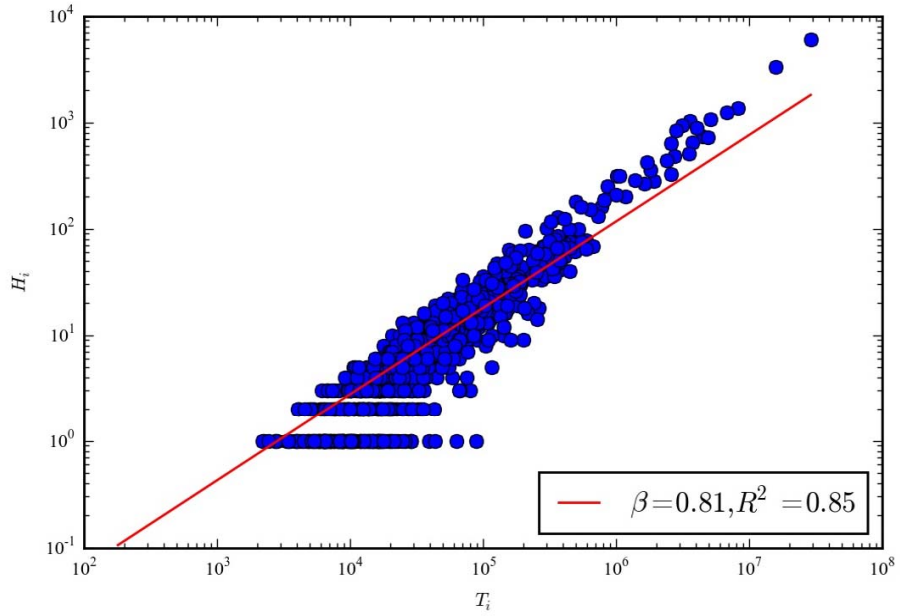

Figure S7 The dissipation law $H_i \sim T_i^{\beta}$ in attention flow network.

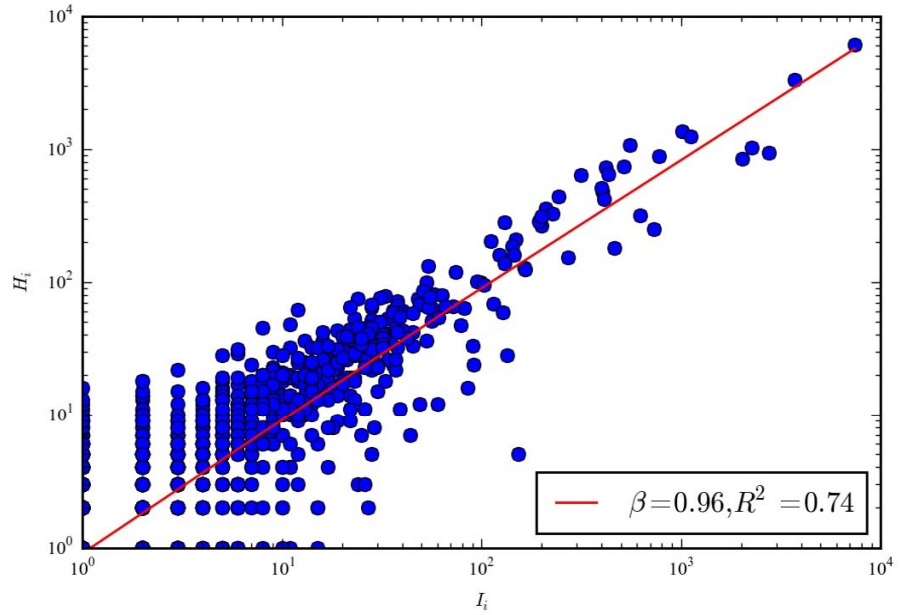

Figure S8 The relationship of H_i and I_i.

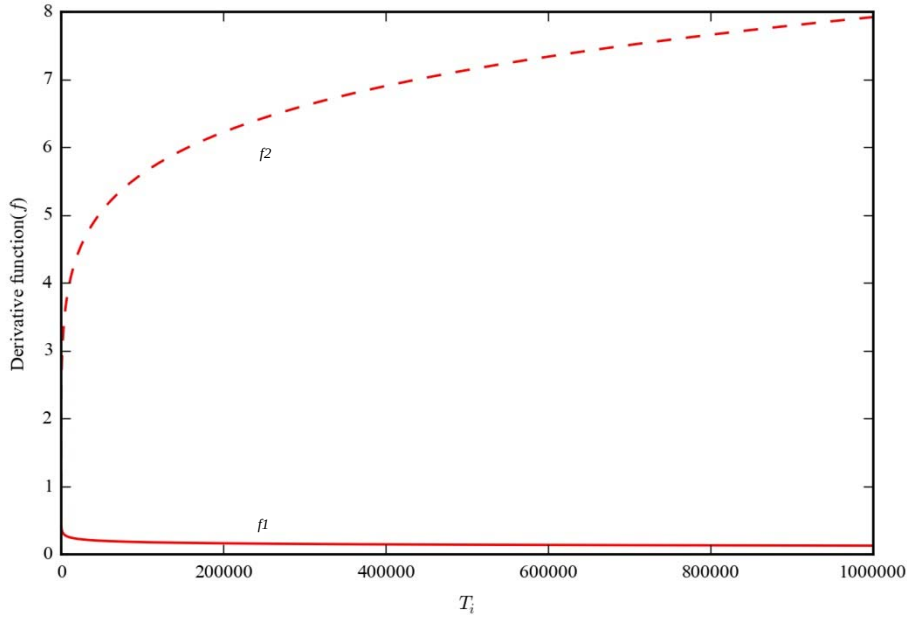

Figure S9 Derivative of the dissipation function.

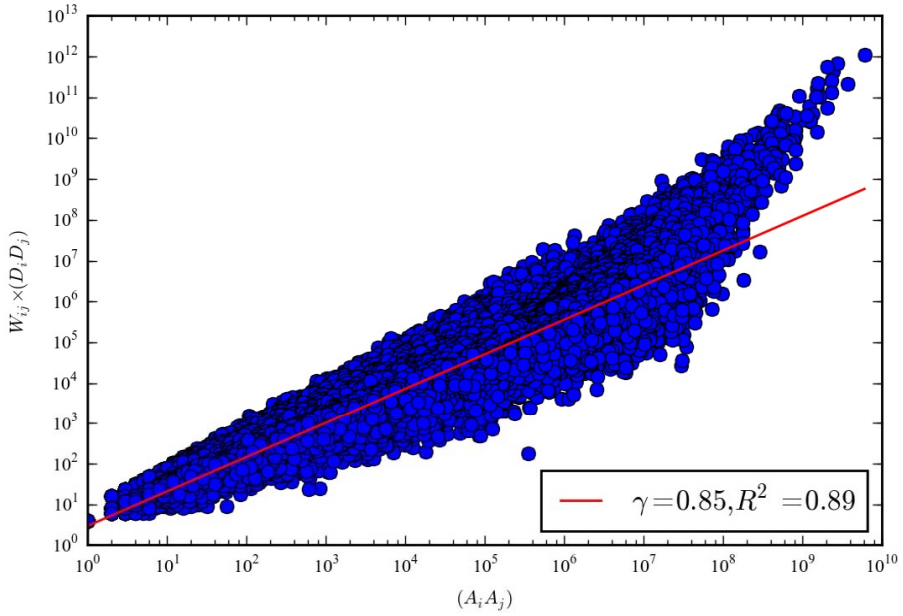

Figure S10 The gravity law in attention flow network.

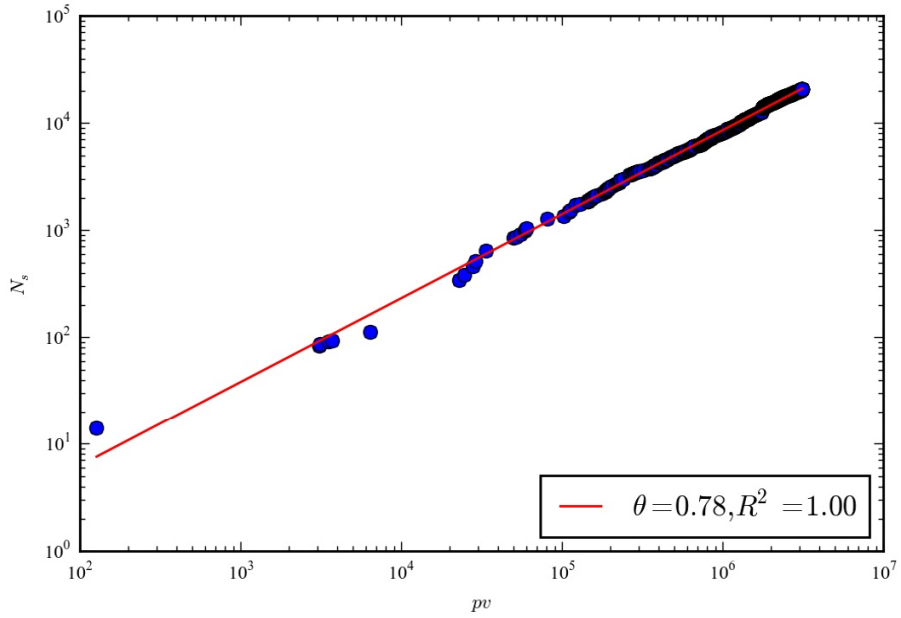

Figure S11 The Heap's law in attention flow network.