OPEN ACCESS Freely available online

PLOS ONE

The Metabolism and Growth of Web Forums

Lingfei Wu[1,2,3], Jiang Zhang[4]*, Min Zhao[1]

1 Baidu Inc., Baidu Campus, Haidian District, Beijing, P. R. China, **2** School of Human Evolution and Social Change, Arizona State University, Tempe, Arizona, United States of America, **3** Center for the Study of Institutional Diversity, Arizona State University, Tempe, Arizona, United States of America, **4** School of Systems Science, Beijing Normal University, Beijing, P. R. China

Abstract

We view web forums as virtual living organisms feeding on user's clicks and investigate how they grow at the expense of clickstreams. We find that PV_t (the number of page views in a given time period) and UV_t (the number of unique visitors in the time period) of the studied forums satisfy the law of the allometric growth, i.e., $PV_t \sim UV_t^{\theta}$. We construct clickstream networks and explain the observed temporal dynamics of networks by the interactions between nodes. We describe the transportation of clickstreams using the function $D_i \sim T_i^{\gamma}$, in which T_i is the total amount of clickstreams passing through node i and D_i is the amount of the clickstreams dissipated from i to the environment. It turns out that γ, an indicator for the efficiency of network dissipation, not only negatively correlates with θ, but also sets the bounds for θ. In particular, $1/\gamma > \theta$ when $0 < \gamma < 1$ and $1/\gamma < \theta$ when $\gamma > 1$. Our findings have practical consequences. For example, θ can be used as a measure of the "stickiness" of forums, which quantifies the stable ability of forums to remain users "lock-in" on the forum. Meanwhile, the correlation between γ and θ provides a method to predict the long-term "stickiness" of forums from the clickstream data in a short time period. Finally, we discuss a random walk model that replicates both of the allometric growth $PV_t \sim UV_t^{\theta}$ and the dissipation function $D_i \sim T_i^{\gamma}$.

Citation: Wu L, Zhang J, Zhao M (2014) The Metabolism and Growth of Web Forums. PLoS ONE 9(8): e102646. doi:10.1371/journal.pone.0102646

Editor: Eduardo G. Altmann, Max Planck Institute for the Physics of Complex Systems, Germany

Received October 15, 2013; **Accepted** June 20, 2014; **Published** August 12, 2014

Funding: This paper is supported by the NNSFC (http://www.nsfc.gov.cn/publish/portal1/) under Grant No. 61004107 and Beijing Higher Education Young Elite Teacher Project (http://www.bjedu.gov.cn) under the Grant No.YETP0291. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors declare that the affiliations to Baidu Inc do not alter their adherence to all the PLOS ONE policies on sharing data and materials.

* Email: zhangjiang@bnu.edu.cn

Introduction

A Web forum is an online discussion site allowing its members to exchange opinions by posting and replying threads. As one of the oldest Internet services, the user-generated-content nature of forums help them thrive in the era of Web 2.0 [1,2]. The popularity of Web forums has motivated various studies on forum-based activities from detecting online opinion leaders [3] and analyzing political debates [4] to identifying interest-groups [5,6]. Due to the challenge of collecting forum browsing data, previous studies usually focus on posting behavior and not browsing behavior. However, the forum usage analysis based on posting dynamics has strong limitations, because there are a large number of "silent" users who only read threads and do not give comments [7,8].

In contrast with the lack of empirical studies on thread browsing, surfing behaviors in other online systems such as tagging sites [9] and social networking sites [7] have been extensively studied. A key concept in surfing dynamics is "clickstream", which either refers to a series of webpages visited in a single session [10], or the successive clicks between two webpages generated by a group of users [11]. Most of early clickstream studies used this term at its first meaning and investigated the distribution of session length l [7,12] and its correlation with other variables, such as session duration [13] and user's log-off probability [14]. In particular, [15] proposed a novel interpretation of the mean value of l as "stickiness", i.e., the ability of a site to keep visitors "lock-in". With the development of network science, there is a trend to use "clickstream" at its second meaning in order to integrate clickstream studies and network theories into clickstream network analysis. In clickstream networks, nodes are information resources and edges are the successive clicks connecting resources [16]. As a general framework, clickstream network has been applied to model various online activities, such as photo tagging [9], news reading [17], and video watching [18]. As demonstrated by these studies, clickstream networks analysis provides novel interpretations to some well-studied problems [19]. For example, the surge and decay of news in the public domain is always understood as a result of the diffusion of information among users [20]. But from the perspective of clickstream networks, it can also be viewed as the transportation of user's attention between news [17].

In the current study we adopt the second definition of clickstream, that is, the successive clicks between two information resources, and use it as a quantification of collective attention online [11]. We get access to the historical data of Baidu Tieba, a very large Chinese Web Forum system, and systematically investigate the browsing activities of users on 30,000 forums in two months. The size (average daily page views) of the studied forums varies from hundreds to millions. We also apply our analysis to two resource sharing forums, Delicous and Flickr, and compare them with Baidu Tieba. Different from previous studies that try to understand how users use forums, we propose to study how forums "consume" user's attention. Specifically, we view forums as "virtual living organisms" that grow at the expense of

user's attention. In this perspective, we discuss the "metabolism" of forums, which describes how the attention of users are "absorbed" into and "dissipated" out from forums. Inspired by the metabolic theory of ecology [21–23], we compare the number of page views as the "body mass" of forums and the number of users as the "energy consumption", and investigate how these two variables are related during the growth of forums. In data analysis, we track the anonymized "cookies", which are permanent, unique identification labels of users, and count the number of unique cookies (UV_t) and page views (PV_t) on an hourly basis. It turns out that the vast majority of the studied forums satisfy the allometric growth law $PV_t \sim UV_t^{\theta}$, which means that the scaling exponent $\theta = d(\log(PV_t))/d(\log(UV_t))$ keeps unchanged over time. We suggest that θ can be used to measure the "stickiness" of forums as an alternative to the average surfing length $L_t = PV_t/UV_t$ [15]. Because both of θ and L_t reflects the ability of forums to remain users "lock-in", but the former is a constant over time, whereas the latter is not.

To probe into the origins of the allometric growth, we construct clickstream networks to define PV_t and UV_t on these networks and explain the observed allometric growth by the interactions between nodes. In particular, we describe the dissipation of clickstreams on nodes using the scaling function $D_i \sim T_i^{\gamma}$ [24,25]. And it turns out that γ, a quantity reflecting the network dissipation efficiency, is negatively correlated with θ. We also conduct a naive mathematical analysis to demonstrate how $1/\gamma$ sets the upper and lower bounds for θ. At the end of our study, we discuss a 2-D random walk model that replicates both of the scaling relationship between PV_t and UV_t and the dissipation function connecting D_i and T_i.

Our study not only confirms the connection between growth and topology in complex systems [21,26–28], but also has applied meanings. For example, the observed universal relationship between PV_t and UV_t will help webmasters to benchmark and monitor the growth of different online communities. Meanwhile, the technique to predict the long-term behavior of forums by analyzing the random snapshots of clickstream networks may contribute to many areas of the Web development, such as click prediction [29] and interest group recommendation; θ as a description of the "stickiness" of forums can be used as a novel feature in the recommendation of interest-groups [30]. Last but not least, we suggest that the presented clickstream network analysis actually provides a very general framework for studying user's browsing behavior in various online systems. To apply our analysis to other types of online social systems, one simply needs to replace the threads (nodes) with other information resources accordingly, such as news, tags, videos, etc.

Materials and Methods

Clickstream networks and key variables

Figure 1 presents an example Baidu Tieba clickstream network, whose nodes are threads and edges are user's switching between threads. The annotation of Figure 1 introduces how to construct clickstream networks from user's log files. We at first divide the entire data set into hourly pieces and then sort each piece by cookies (the unique and permanent labels used by a website to identify users). After that, we select all successive pairs of threads visited by the same user and connect them in the clickstream network. Sorting data by cookies guarantees that a user would not be repeatedly counted even if he is logged in/out more than once during a hour, so UV_t always represents the unique number of users.

We find that, after we adding "source" and "sink" to balance them [43], clickstream networks satisfy the constrain of "clickstream conservation". Thus, PV_t and UV_t as network properties, can also be calculated at the node level. As network properties, PV_t is the total weights of edges and UV_t is the total clickstreams "dissipated" out of the network (i.e., the weighted in-degree of "sink". Note that UV_t also equals the weighted out-degree of "source", thus we can choose either "source" or "sink" to conduct the analysis. To make our clickstream networks comparable with ecological networks [24], we choose to analyze "sink". See Figure S3 for the comparison between the dissipation behaviors calculated by "source" and "sink"). On the node level, PV_t is the sum of the clickstreams passing through node i (T_i) and UV_t is the sum of the clickstreams dissipated by i to "sink" (D_i):

$$PV_t = \sum_i T_i, \tag{1}$$

$$UV_t = \sum_i D_i. \tag{2}$$

Data

Two groups of data sets are used. The first one is the log file of Baidu Tieba (http://tieba.baidu.com/), a collection of many topic-specific forums. Among the millions of forums in the system, we select the top 30,000 forums, whose size (the averaged daily page views in two months) varies from hundreds to millions. For each forum, we construct 1,440 successive hourly-based clickstream networks using the historical browsing data in two months (from Feb. 27, 2013 to Apr. 27, 2013). The other group of data sets contains the historical log file of two popular tagging sites, Delicious (https://delicious.com) and Flickr (http://www.flickr.com). These two data sets are collected by the joint effort of the institutions in the TAGora European project (http://www.tagora-project.eu/data/), which have generated many papers including [9] and [38]. The Delicious data set covers individual tagging behavior in four years (from 2003-01-01 to 2006-12-28) and the Flickr data set covers tagging behavior in two years (from 2004-01-01 to 2005-12-31).

In constructing Flickr and Delicious clickstream networks, we use the same method as illustrated in Figure 1, except that the nodes (which were threads in Baidu networks) are now the tags used by users to annotate online resources and the links are the successive usage of two distinct tags. Meanwhile, although Tieba networks are constructed in an hourly basis, we construct Flickr and Delicious networks in a daily basis so that they all contain $10^2 \sim 10^4$ nodes and thus are comparable in size (see Figure S1). Despite these differences, our analysis shows that both types of clickstream networks exhibit very similar behaviors. Due to the data usage constraints, we are not able to release Tieba data. But we provide the download of Delicious and Flickr daily clickstream networks in http://pan.baidu.com/s/14Csma and http://pan.baidu.com/s/1gdsWMSN, respectively.

Results

The allometric growth of forums

Kleiber's law, or allometric growth, predicts that for a majority of living organisms, their energy consumption scales to body size

A

a	0
a	1
a	2
b	0
b	2
c	2
c	3
d	0
d	4
d	5
e	5
e	0

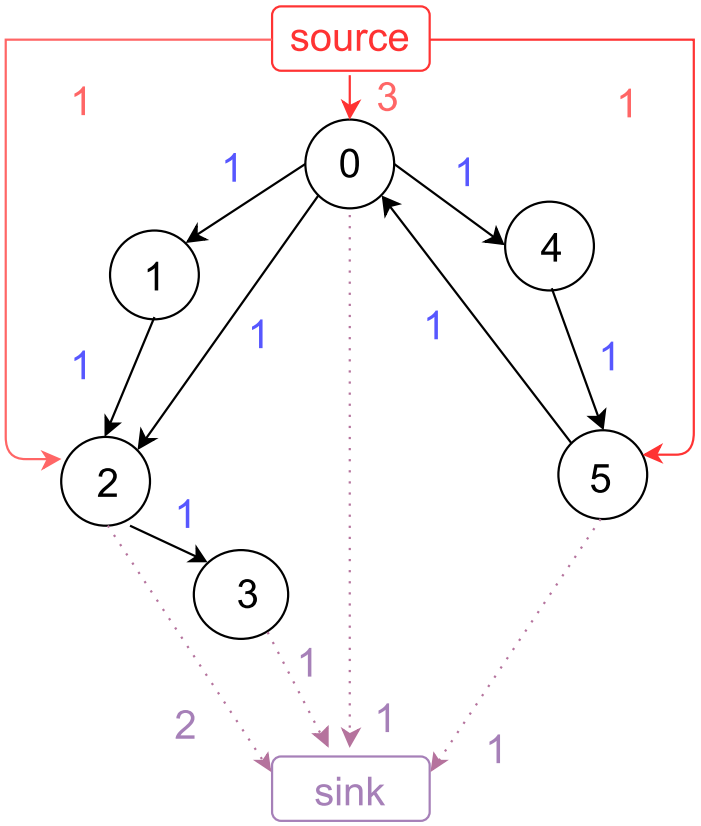

Figure 1. An example dataset of Baidu Tieba log file in one hour and the corresponding clickstream network. In (A) the left column shows the anonymized, sorted cookies and the right column shows the numeric ID of the visited threads. In (B) the nodes are threads and the weighted, directed links are user's switching between threads. The red arrows show the clickstreams "absorbed" from "source" and the purple, dotted arrows show the clickstreams "dissipated" to "sink". In particular, the network in (B) is constructed as follows. For each record in the dataset, say, [a, 0], if the next record has the same cookie, e.g., [a, 1], we add a clickstream from node 0 to node 1; otherwise, we create a clickstream from node 0 to the artificially added node "sink". After all records are converted into clickstreams, we add a "source" node to balance the network such that in-flow (weighted in-degree) equals out-flow (weighted out-degree) over all nodes except "source" and "sink" [43]. In the constructed networks, the values of passing-through clickstreams T_i from node 0 to 5 are $\{4,1,3,1,1,2\}$, and the values of corresponding dissipation D_i are $\{1,0,2,1,0,1\}$. The values of PV_t and UV_t of this network are 12 and 5, respectively. Note that the value of PV_t equals the total number of records in (A) and also the sum of T_i, and the value of UV_t equals the total number of users in (A) and also the sum of D_i.
doi:10.1371/journal.pone.0102646.g001

with an exponent equals 3/4 [21]. If we view online communities as virtual living organisms that feed on user's attention, a particularly interesting question would be, what are the counterparts of "body mass" and "energy consumption" of these virtual entities? Banavar et al. [27] explain Kleiber's law by modeling living organisms as flow networks that transport waters and nutrient. In their model, "body mass" is the total amount of flow circulating within a network and "energy consumption" is the amount of flow the network exchanges with the environment. By applying this model to clickstream networks, one would immediately find that these are also the definitions of "PV" (the total number of page views or clicks in a given period) and "UV" (the total number of unique user sessions in the given period) of websites, respectively. Therefore, the online version of Kleiber's law, to exist, predicts that,

$$PV_t = aUV_t^{\theta}, \tag{3}$$

in which a is a constant coefficient. The exponent θ in Eq.3 not only shapes the growth dynamics of forums, but also provides a measure of the "stickiness" of forums as an alternative to the average surfing length L, which is suggested in [15]. Using the indicator of θ, we can easily separate "sticky" forums from "non-sticky" forums. In particular, we derive that

$$L_t = PV_t/UV_t \sim PV_t^{1-1/\theta}. \tag{4}$$

If $\theta > 1$ and hence $1-1/\theta > 0$, the average surfing length of users increases with forum size (or "body mass"). In other words, users are more likely to be "locked-in" in a forum during its growth. This is what we expect to see from a "sticky" forum. On the contrary, if $\theta < 1$ and hence $1-1/\theta > 0$, users on average navigate less threads as the size of the forum increases, which is the property of a "non-sticky" forum. An extra bonus of using θ as the indicator is that, $\theta = d(\log(PV_t))/d(\log(UV_t))$ is a constant over time, whereas L_t is obviously not. Therefore, θ quantifies the "stickiness" of forums as a stable, long-term property.

Figure 2 demonstrates that Eq.3 characterizes the growth dynamics of three different forums and two tagging systems during the studied period. We find that this strong regularity holds for most of the studied forums: more than 86% of forums have $R^2 > 0.8$ in the fitting of Eq.3. This finding suggests that the users of different forums obey similar behavioral logic in browsing

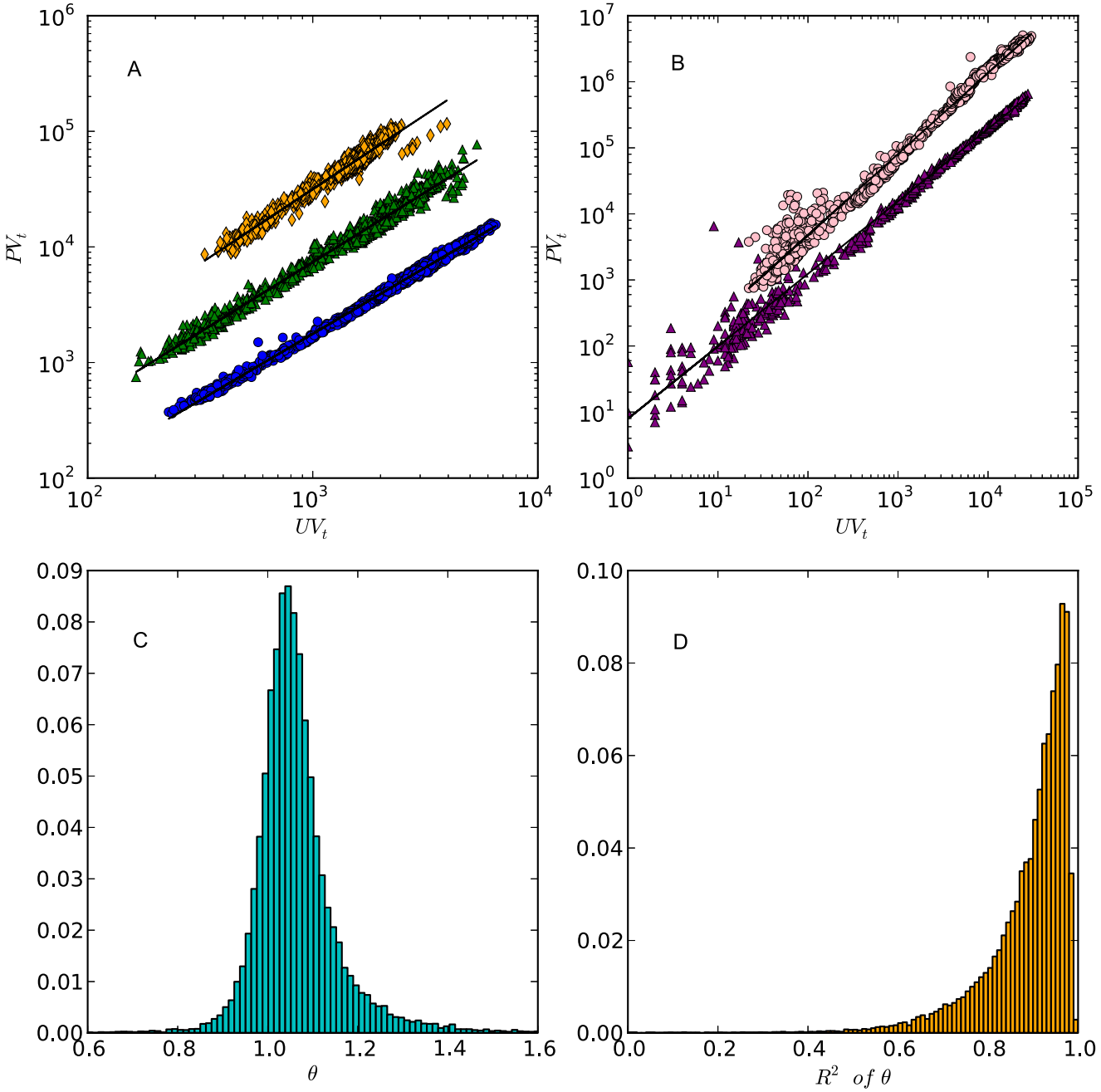

Figure 2. The scalings between UV_t and PV_t across three forums in 1,440 hours, (A). Each data point corresponds to a pair of UV_t and PV_t for an hourly network. Data points of different forums are shown in different colors. The values of θ are 1.15 (blue circles), 1.21 (green triangles), and 1.29 (orange diamonds), respectively. (B) The scalings between UV_t and PV_t of Delicious (pink circles) and Flickr (purple triangles). Each data point corresponds to a pair of daily UV_t and PV_t. The values of θ are 1.23 and 0.10, respectively. (C) The distribution of θ of 29,993 forums (the estimation of the rest 7 forums are removed due to a lack of data). The mean value is 1.06 and the standard deviation (SD) is 0.10. (D) The distribution of R^2 in fitting the θ of Baidu forums. The mean value is 0.89 and the SD is 0.10.
doi:10.1371/journal.pone.0102646.g002

threads collectively. It is very inspiring to find that human attention, after being quantified as clickstreams, satisfies the physical laws observed widely in natural flow systems [31].

In Kleiber's law, the "body mass" scales to "energy consumption" with an exponent $4/3 \approx 1.33$ [21]. But the exponent observed in our data is generally smaller than this value. We conduct KS test [32] to verify the assumption that the calculated θ is drawn from a normal distribution with a mean equals 1.06 and a standard deviation (SD) equals 0.10. The p-value of the KS test is 0.07, suggesting that we can not reject this assumption. As shown by Figure 2C, the shape of the distribution is slightly asymmetrical; it skews towards the right hand side of the x axis beyond the point of ($x = 1$, $y = 0$). In fact, 82% of the forums have a $\theta > 1$. Thus most of the studied forums are "sticky", in the sense that users are more likely to remain in the forums when the forums grow in size. However, by comparing θ between virtual and real

flow systems, we find that clickstream networks are still not as "sticky" as energy transportation networks within living organisms [21]. How can websites learn from living organisms? This is an interesting topic worth further exploration.

The scaling of clickstream dissipation

We also discover an interesting scaling between T_i and D_i that describes the dissipation behavior of nodes,

$$D_i = bT_i^{\gamma}, \tag{5}$$

in which b is a coefficient and γ is an exponent that reflects the efficiency of network dissipation.

To understand the meaning of γ, we can define the log-out probability of users on node i as

$$P_i = D_i/T_i. \tag{6}$$

Thus, $P_i \sim T_i^{\gamma-1}$ if Eq.5 holds. P_i increases with the clickstreams passing-through nodes if $\gamma > 1$, and decreases with the clickstreams otherwise. Therefore, the dissipation efficiency γ quantifies how the log-off probability changes with the node traffic T_i.

Although γ in Eq.5 seems to depend heavily on the flow structure of clickstreams networks, which may change in time, it is actually very stable during the growth of clickstream networks. We randomly select a day (Apr. 24, 2013) and construct 24 successive hourly networks for each of the studied 30,000 forums. We find that to estimate γ, we just need one hourly clickstream network. The values of γ estimated from 24 networks have a very small standard deviation (SD). Figure 3 shows that more than 98% of forums have an $R^2 > 0.8$ in the fitting of Eq.5. Meanwhile, the value of γ estimated from hourly networks is a stable quantity over time (the SD of γs in 24 hours is 0.14). We conduct KS test to verify the assumption that the calculated mean value of γ is drawn from a normal distribution with a mean equals 0.93 and an SD equals 0.08. The p-value of the KS test is 0.14, suggesting that we can not reject this assumption. The distribution of γ skews towards the left hand side of the x axis beyond the point of (x = 1, y = 0) and 82% of forums have a value of γ smaller than 1. According to aforementioned discussions, this means that most of the studied forums have a low dissipative efficiency, i.e., the log-out probability of users decreases with the clickstreams passing through threads.

This finding provides insight into the usage of Tieba forums by implying that users are more likely to log out from non-popular threads than popular threads. This is because Tieba system sorts threads in the reversed chronological order of comments and displays threads in multiple pages. Therefore, popular threads who receive more comments always appear on the first page. Unlike News aggregators such as Yahoo!, Tieba is an interested-based community containing topic-specific forums, therefore instead of selective reading, users usually simply browse the threads one by one in the default displaying order. As a result, when users get tired, they usually have read the most popular threads.

The negative correlation between γ and θ

By reviewing Eq.3 $\sim$ Eq.5, one would naturally expect that the dissipation efficiency γ and the stickiness θ are related. To understand the connection between the two parameters, let's consider two extreme topologies, the star-like (Figure 4A) and the chain-like (Figure 4B$\sim$C). In the star-like topology, threads (nodes) receive clickstreams directly from the "environment" and dissipate them immediately, whereas in the chain-like topology, threads transport clickstreams from one to another and dissipate a portion of clickstreams in each step. If we fix the UV_t of the three clickstream networks to be the same as 10 units, we will find that the resulting PV_t is different: it is larger in the chain-like networks (10+3+1.5+1+0.9 = 16.4 in B and 10+9+6+3+0.9 = 28.9 in C) than in the star-like network (3+2.5+1.5+1 = 10 in A). This is because by transporting clickstreams a network increases its storage capacity of clickstreams, i.e., the "body mass".

To understand this interesting phenomenon, one can consider how a clown plays balls. A clown can barely hold more than two balls if he just grasps them in his hands, but he can easily maintain a circulation of many balls by throwing them up and passing them between hands. It is in exactly the same way that clickstream transportation increases the total amount of clickstreams "hold" by a network.

Furthermore, we find that a small γ would decrease the dissipation of clickstreams and thus increases the network storage capacity. This finding is demonstrated by the comparison between Figure 4B and C. We calculate that $P_{ib} = \{70\%,50\%,30\%,10\%\}$ from node A to D in Figure 4B and $P_{ic} = \{10\%,30\%,50\%,70\%\}$ in Figure 4C (for the convenience of the comparison, we ignore the behavior of node E, whose traffic is very small compared to other nodes). As the pass-through clickstreams decrease monotonously from A to D, it is easy to derive that $\gamma_b > 1 > \gamma_c$. Recalling the conclusions that $UV_b = UV_c$ and $PV_b < PV_c$, which imply that $\theta_b < \theta_c$, we find that γ and θ are negatively correlated. In fact, it is reasonable to expect this negative correlation being applicable to clickstream networks of all kinds of topologies. Because a small γ will always force large nodes to transport clickstreams to other nodes rather than dissipating them to the environment.

Figure 5A shows that the empirical data support the negative correlation between γ and θ. To summarize, the reversed chronological displaying order of threads seems to decrease the dissipation efficiency γ and increase the "stickiness" θ of the studied forums. This may be the reason why such displaying order is so common among forums. The web masters may or may not have noticed that, this strategy beats its competitors by generating a flow structure that attracts more users and thus spreads out in the evolution of forums.

As a complementary analysis, we also examine whether γ and θ are affected by the forum size. We plot these two quantities against forum size in Figure 5B and find that when the forum size approximates 10^5 daily views, γ reaches its minimum value and θ reaches its maximum value. This observation can be used to benchmark the growth of Tieba forums.

$1/\gamma$ as the Bound of θ

Negative correlation is not the only connection between γ and θ. Here we present some derivations to demonstrate that γ actually sets the bounds for θ. We can put Eq. 3, Eq. 1, Eq. 2, and Eq. 5 together as

$$\begin{cases} PV_t = aUV_t^{\theta} \\ D_i = bT_i^{\gamma} \\ PV_t = \sum_1^k T_i \\ UV_t = \sum_1^k D_i, \end{cases} \tag{7}$$

in which $T_i > 0$ and $D_i > 0$.

If $\gamma > 1$, then $1/\gamma < 1$. Assuming that there are k nodes in the network, we can derive that (see SI for the derivation in details)

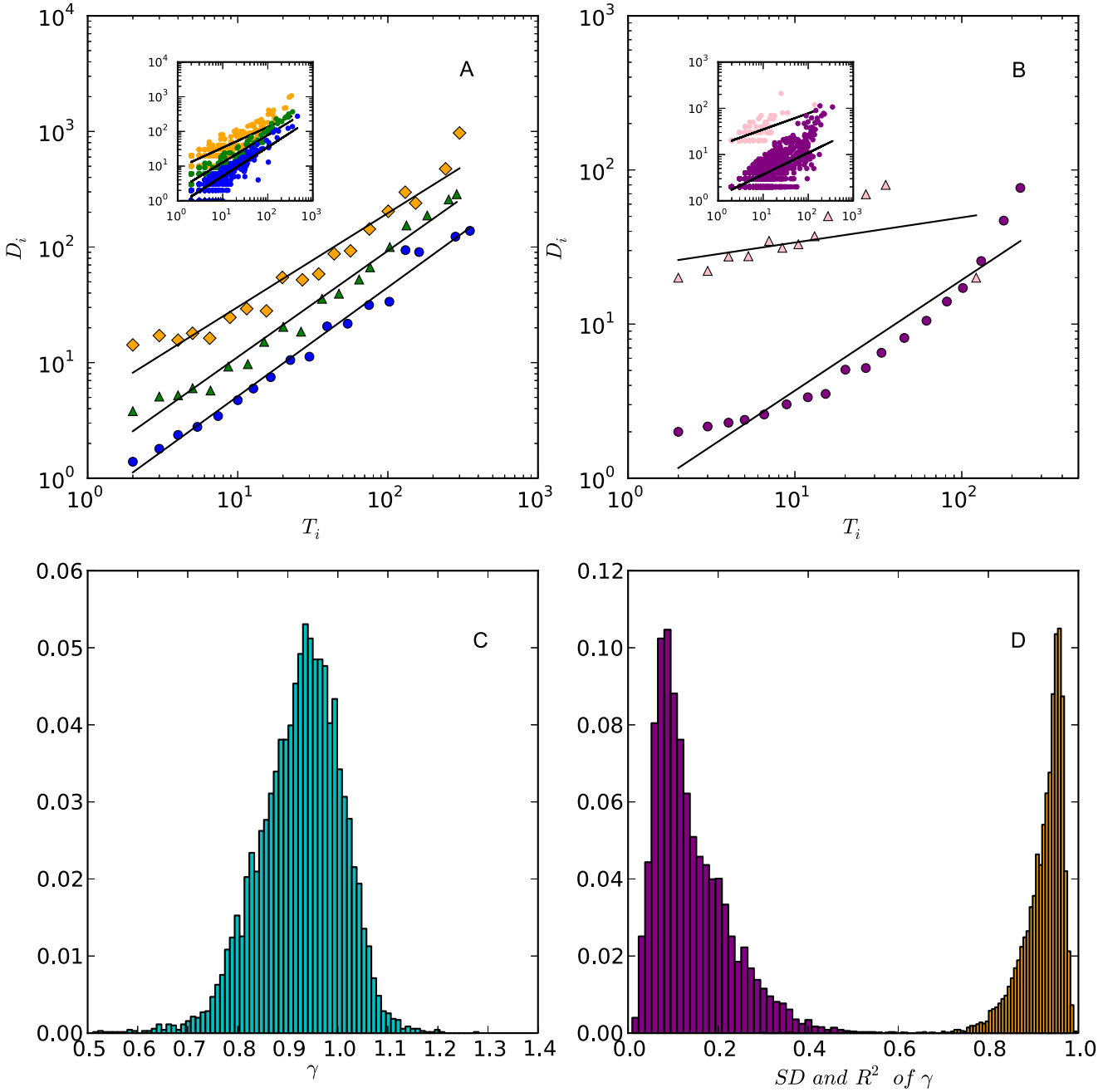

Figure 3. The scalings between T_i and D_i across forums in three hourly networks, (A). These three forums are the same as the forums presented in Figure 2A. The color scheme of these data points is the same as that of Figure 2A. The value of γ are 0.96 (blue circles), 0.90 (green tringles), and 0.80 (orange diamonds) for the three forums, respectively. (B) The scalings between T_i and D_i of Delicious (pink triangles) and Flickr (purple circles) in 2004-12-01. The values of γ are 0.83 (Flickr) and 0.29 (Delicious), respectively. In both of (A) and (B), the regression estimation is applied on the log-binned data, in which we calculated the average of x and y values in the intervals uniformly selected from the e-based logarithmic scaled x range. This technique is frequently used to eliminate the noise in data [44]. We also present the raw data in insets. (C) The distribution of the averaged value of γ over 24 hours across $6{,}877$ forums in Baidu Tieba. The mean value of the distribution is 0.93 and the SD is 0.08. (D) The distribution of the SD of γ over 24 hours (purple bars) and the averaged R^2 in fitting γ (orange bars). The mean and SD of the two distributions are 0.14, 0.09, and 0.92,0.05, respectively.
doi:10.1371/journal.pone.0102646.g003

$$aUV_t^{\theta} = PV_t = \sum_1^k T_i = (\frac{1}{b})^{1/\gamma} \sum_1^k D_i^{1/\gamma} > (\frac{1}{b})^{1/\gamma} (\sum_1^k D_i)^{1/\gamma} = (\frac{1}{b})^{1/\gamma} UV_t^{1/\gamma}, \quad (8)$$

and that

$$aUV_t{}^{\theta} = (\frac{1}{b})^{1/\gamma} \sum_1^k D_i^{1/\gamma} < (\frac{1}{b})^{1/\gamma} \sum_1^k D_i = (\frac{1}{b})^{1/\gamma} UV_t. \quad (9)$$

Putting Eq. 8 and Eq. 9 together we have

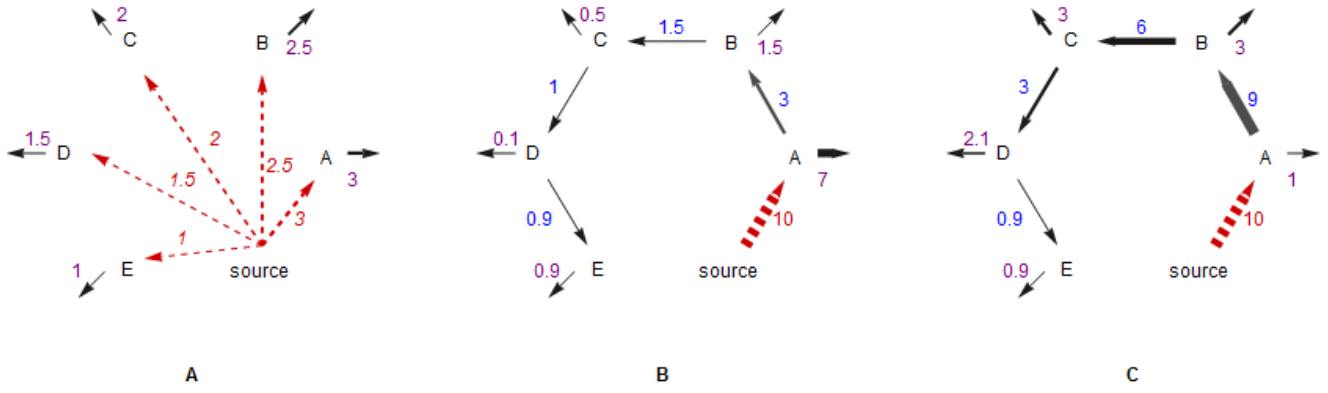

Figure 4. Three example clickstream networks of different topologies. The dashed red arrows show the clickstreams from source to the other nodes in the network. The black arrows show the transportation of clickstreams within the networks (the weights of which are shown in blue letters) and the dissipation of the clickstreams out of the networks (the weights of which are shown in purple letters). (A) A star-like network in which the dissipation probability of all nodes equals 100%. (B) A chain-like network in which the dissipation probability $P_i = \{70\%,50\%,30\%,10\%\}$ decreases from node A to D. As the clickstreams passing though nodes (T_i) also decreases from A to D, P_i is positively correlated with T_i. (C) A chain-like network in which the dissipation probability $P_i = \{10\%,30\%,50\%,70\%\}$ increases from node A to D. P_i is negatively correlated with T_i. According to Eq.5 and Eq.6, we are able to derive that $\gamma_b > 1 > \gamma_c$. As we also know that $UV_b = UV_c$ and $PV_b < PV_c$, which implies that $\theta_b < \theta_c$, we find that γ and θ are negatively correlated. In the above naive comparison, we ignore the behavior of node E, whose traffic is very small compared to the other nodes.
doi:10.1371/journal.pone.0102646.g004

$$UV_t^{1/\gamma} < ab^{1/\gamma} UV_t^{\theta} < UV_t. \quad (10)$$

Compare to UV_t, whose value varies from 10^2 to 10^4 (Figure 2), the value of $ab^{1/\gamma}$, which varies from $1 \sim 2$ (see Figure S2), is very small. Therefore, the following inequality should be satisfied to guarantee Eq. 10:

$$1/\gamma < \theta < 1 \ (\gamma > 1). \quad (11)$$

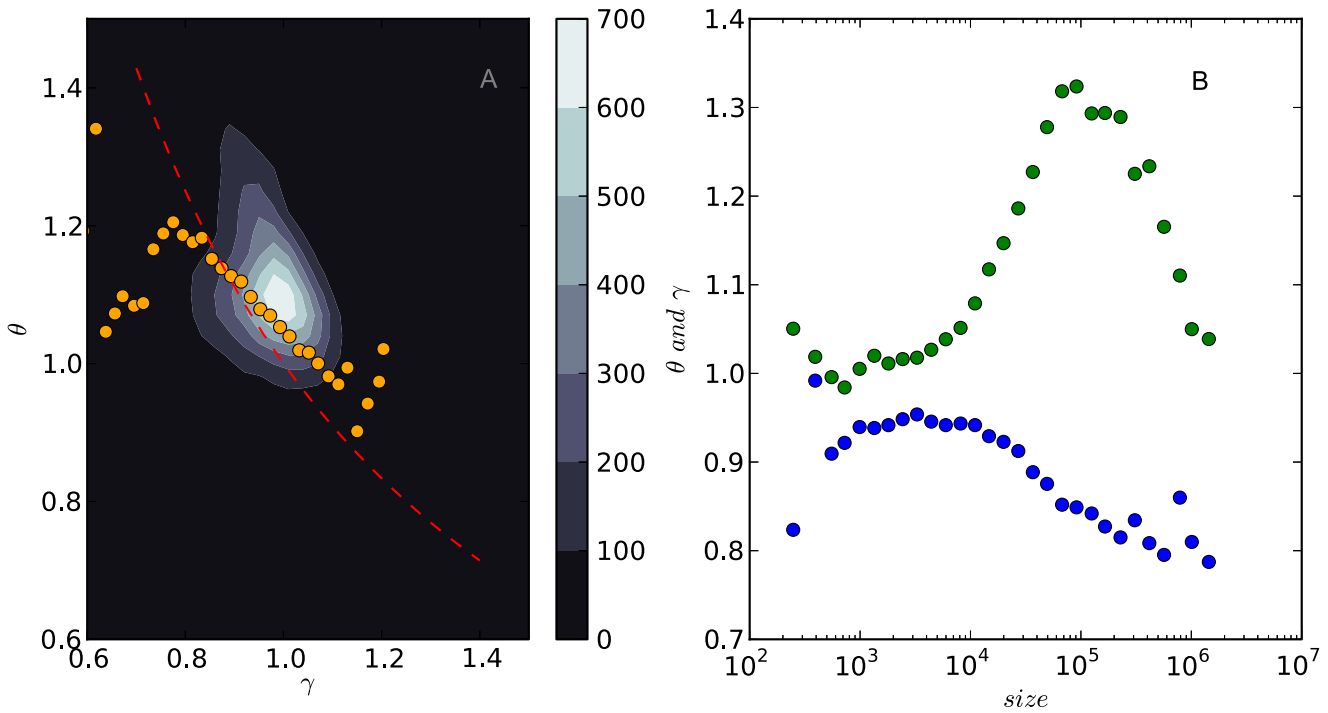

Figure 5. The negative correlation between γ and θ (A) and the change of γ (blue circles) and θ (green circles) with forum size (B). In (A) We plot both of the linear-binned data (orange circles) and the original data (heat map) and in (B) we only show the linear-binned data. In the heat map, the lighter color means that the distribution of the data points is more dense. The ticks on the color bar show the number of data points within a 0.1*0.1 square.
doi:10.1371/journal.pone.0102646.g005

Similarly, when $0<\gamma<1$ and thus $1/\gamma>1$ we can derive that

$$1<\theta<1/\gamma\ (0<\gamma<1). \quad (12)$$

We find that Eq. 11 and Eq. 12 are supported by Figure 5. When $0<\gamma<1$, $1/\gamma$ (the red, dotted line) is the upper bound of the expect values of θ (the orange circles); when $\gamma>1$, $1/\gamma$ becomes the lower bound.

Discussion and Conclusion

A model of individual surfing behavior

We propose a simple model that replicates the observed two scaling laws (Eq. 3 and Eq. 5). Two properties of surfing behavior feature our model: 1) users can both browse existed threads and also publish new threads; 2) the previous surfing activities have an effect on the following surfing activities.

Specifically, we model surfing activities by random walks within a 2-D grid containing randomly distributed threads. The Euclidean distances between threads indicate their similarities. To initialize the simulation we place only one thread (the red point in Figure 6 A) as a seed at the center of the universe. In each iteration we drop an fixed number of users uniformly to the system, who will walk randomly on the grid and create new threads with probability p until there are no existed threads within their "observation zones" (an $r\times r$ square around their current position). This is to represent that users will leave the forum if they can not find interested threads within a search area. We place no constraint on user's random walk so a thread may be visited repeatedly. But if a user can not find existed threads within the "observation zone" at the first step of his random walk, he will leave the system immediately and does not contribute to the statistics mentioned in the following part.

A random walk in this model represents the browsing activities generated by a user in a given time period. Therefore, the number of random walks can be expressed as UV_t and the total number of repeatedly visited threads is PV_t. As time goes by, threads are created and are connected by user's random walks, leading to a growing clickstream network on the grid (Figure 6 B), which attracts more users and allows longer random walks. If we define T_i as the total number of random walks visiting thread i and D_i as the number of particles leaving the system from i, we will find that Eq. 1 and Eq. 2 still hold. This is because there is also "clickstream conservation" in this model; the number of users entering into the system equals the sum of users leaving the system over all nodes, and the total number of repeatedly visited threads equals the sum of visits to each thread. As shown by Figure 6 C and D, our model demonstrated both of the allometric growth (Eq. 3) and the scaling law of dissipation (Eq. 5). Our model also exhibits the negative correlation between γ and θ (see Figure S4), although this relationship is not significant (Pearson correlation coefficient equals -0.23 and p-value equals 0.5).

We conjecture that, the observed super-linear scaling between PV_t and UV_t in our model originates from the fractal flow network structures [26] at the early stage of the simulation. A strong limitation of our model is that, as time passes, this fractal structure converges to a completely filled 2-D disk. This explains why the scaling exponent γ evolves towards 2 (in theory, a random walker can visit any point within a 2-D space, so the average length $L_t\sim UV_t$ and hence $PV_t=L_t*UV_t\sim UV_t^2$) and also why the dissipation exponent γ evolves towards 0 (eventually the dissipation only happens on the edge of the disk, so the average dissipation of all threads on disk approaches 0).

The novel feature of this model is that it demonstrates how flow creates a structure that attracts more flow. The co-evolution between structure and flow makes this model very different from previous network models, which either focus on the dynamics of networks [33,34] or the dynamics on networks [20], but not both of them.

The distribution of forum categories in the $\gamma\times\theta$ space

Driven by practical interests, we investigate whether the content of forums relates to their stickiness and dissipation efficiency. Figure 7 gives the distribution of 22 categories of 9,978 forums (the rest of the top 30,000 forums are removed due to a lack of human labeling data) in the $\gamma\times\theta$ space. Each circle corresponds to a category of forums labeled by human coders. The size of green and orange circles reflects the average size and the number of forums in the corresponding category, respectively. We observe that γ and θ are negatively correlated, which is consistent with the findings in Figure 5.

This analysis provides insights for the management of Tieba forums. For example, the categories "Art", "Cartoon", and "Personal space" locate at the lower-right corner of the space, suffering from high dissipation efficiency and low stickiness. It means that on these forums users do not read a lot of threads within a single session. On the contrary, the categories "Beijing Olympics", "female", and "sports" have high stickiness, suggesting that users to these forums are generating more clicks. In particular, the high value of θ of the "Female" category suggests that there are a lot of female Tieba users. This conclusion is supported by the user statistics of Alexa (www.alexa.com), which suggests that the proportion of female users of Baidu Tieba is higher than the average level of the Web users.

Summary

Websites, by their very nature, are the consumers of collective attention and the producers of information [35]. The comparison of websites as living organisms is not just a qualitative metaphor, but also provides quantitative insights into the understanding of websites development. In this study, we find substantial evidence that the growth dynamics of websites is governed by laws that are known to shape the evolution of natural flow systems [21].

In particular, we discuss the online version of Kleibers' law, that is, the scaling between UV_t and PV_t in the temporal evolution of forums. Furthermore, we show that the allometric exponent θ, which is an indicator for the "stickiness" of forums in attracting users, is determined by the metabolism of clickstream networks. The lower the dissipation efficiency γ is, the larger the θ would be. Interestingly, there seems to be an optimized scale of forums at around 10^5 daily PVs that minimizes θ and maximizes γ. Finally, we discussed a random walk model that replicates both of the allometric growth and the dissipation patterns.

As suggested by Bettencourt et al. [36], the allometric growth is a very general relationship between variables in the evolution of complex systems. In particular, they show that cities are extensions of biological entities, in the sense that they satisfy the same allometric functions [22,36]. Our study extends their findings from offline social systems to online social systems. We are not the only researchers who have noticed the scaling laws in online communities. For example, the recently found "densification" pattern in the growth of online networks [37], together with the scalings discussed in [37–41], are different versions of the "allometric growth" of online flow networks.

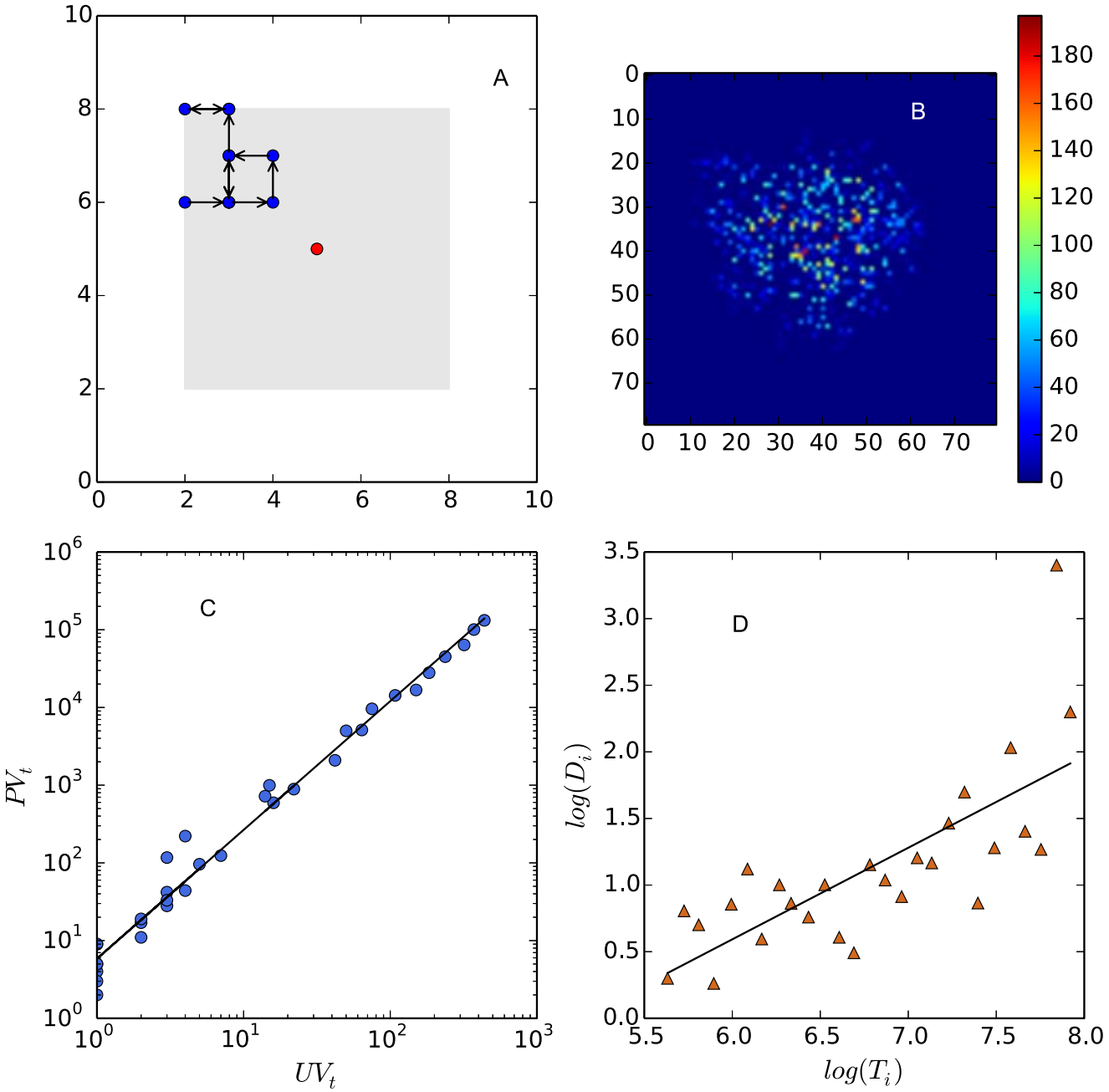

Figure 6. A model of individual surfing behavior. (A) The arrows show the trajectory of a random walker who starts from (2, 6) and ends in (3, 8). New threads (blue points) are created with probability 1 by the random walks. The walker stops when he goes beyond the "observation zone" (the gray square) of the existed threads generated in the last iteration of simulation. To initialize the simulation, we place a seed thread in the center of the grid as the existed thread (the red point). (B) The 1,000 generated threads (85 iterations, $r=3$, $p=0.05$) within a $1{,}000\times1{,}000$ space. At the beginning of the simulation, there is only one thread at the center of the space. For each iteration, we throw $1{,}000$ particles uniformly distributed into the space. We use warmer colors to show the larger number of visits to threads. In this plot we only show the central part of the original space in order to obtain a higher resolution network structure. (C) The increase of the total number of repeatedly visited threads (PV_t) with the number of user sessions (random walks) (UV_t). Each data point corresponds to an iteration in the simulation. The scaling exponent θ is 1.63. (D) The increase of the number of particles leaving the system from thread i (D_i) with the number of total visits to i (T_i), both axes are shown in e-based logarithmic scale. The exponent γ is 0.77.
doi:10.1371/journal.pone.0102646.g006

Our findings are relevant to the Web development in many aspects. In particular, the presented method predicts the long-term trend of clicks thus is useful in computational advertisement [42]. To predict the "stickiness" θ of forums, one just need to collect a random sample of threads and record the clickstreams passing through and being dissipated by them in a single hour. Another possible application of θ is to use it as a novel feature in the recommendation of interest-based groups [30].

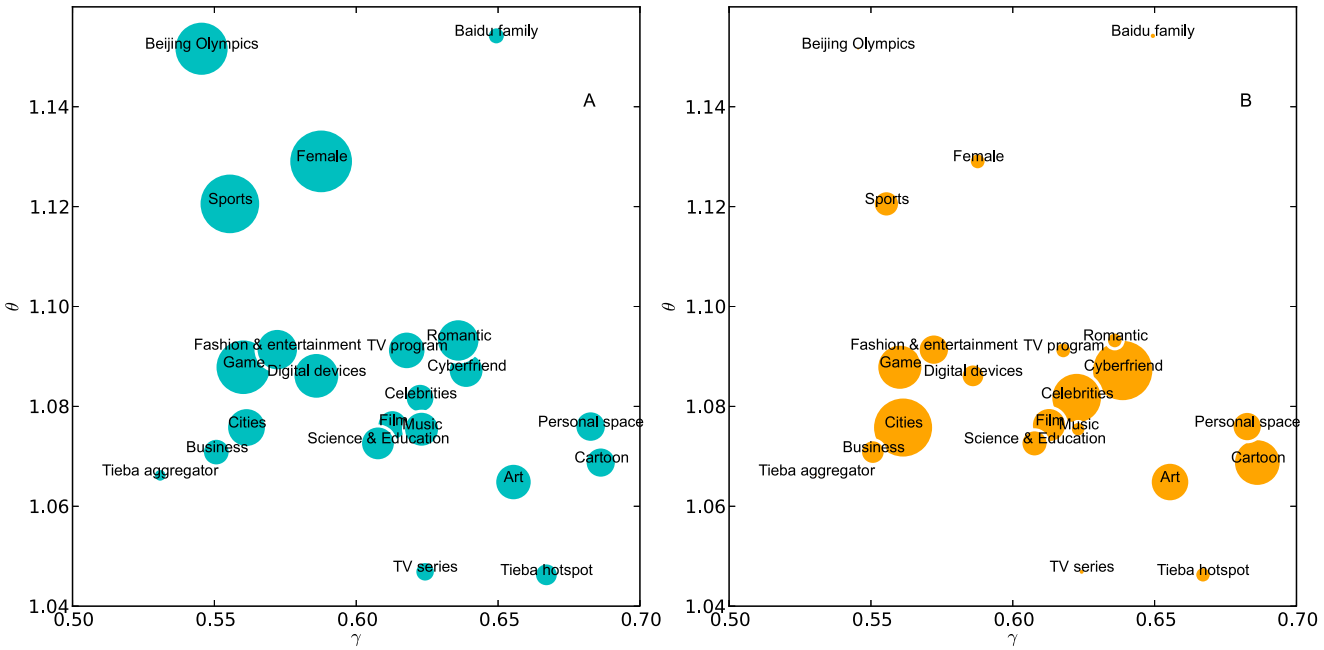

Figure 7. The mean values of γ and θ of different categories of forums of Baidu Tieba. Each circle corresponds to a category of forum. The size of green circles in (A) reflects the average size of forums in the category and the size of orange circles in (B) shows the number of forums within the category.
doi:10.1371/journal.pone.0102646.g007

Supporting Information

Figure S1 Three snapshots of clickstream networks of Delicious. (A), (B), and (C) show the networks in 2003-01-01, 2003-06-01, and 2003-12-01, respectively. In each network, the nodes are tags and the weighted links are the sequential usage of two tags by users. In these networks source and sink are denoted by green and red colors, respectively. Other nodes are clustered by the week components they belong to and the nodes from the same cluster are shown in the same color. The networks are connected by source and sink and will fall apart if we remove these two nodes. It is observed that as the networks evolve, the largest component (in blue color) grows, connecting frequently used tags.
(EPS)

Figure S2 The empirical distributions of the parameters of Eq. 8. The value of a of each forum is estimated from the scaling relationship between UV_t and PV_t (see Eq. 8) in 1440 hours. In estimating the values of b we construct 1440 hourly flow networks for each forum, estimate the hourly scaling exponent between T_i and D_i (also see Eq. 8), and calculate the mean of the hourly values. The distributions shows the parameters for the top 1,000 forums.
(EPS)

Figure S3 The linear relationship between γ_D and γ_I. We plot both of the \binned" data (orange circles) and the original data (heat map). In the heat map, the lighter color means that the distribution of the data points is more dense. The slope of the regression line fitted from the binned data is 0:46.
(EPS)

Figure S4 Some results of simulation. (A) The change of θ (blue points) and γ (red points) with the thread generating probability p. (B) The negative correlation between θ and γ. The Pearson correlation between the θ and γ is -0.23, which is consistent with the empirical findings. However, this estimation has a p-value equals 0.5, thus we fail to significantly rule out the probability that the two parameters are independent. Simulations on the larger scales are needed to conform the relationship between θ and γ in this model.
(EPS)

Acknowledgments

The authors thanks Xingyuan Yuan for his help in setting up the interactive clickstream network visualization on Swarm Agent Club's server. The authors also thanks Dejun Yu, Bo Yang and other people who attended the "Complex Network and Flow" seminar series. L. Wu thanks Baidu colleagues Kaiyuan Fang, Qiwen Liu, and Jiangyun Song for stimulating discussions and Tianjian Chen and Junying Zhang for supporting Wu's internship and research in Baidu Inc.

Author Contributions

Conceived and designed the experiments: JZ. Performed the experiments: LW MZ. Analyzed the data: LW. Contributed reagents/materials/analysis tools: MZ. Wrote the paper: LW JZ.

References

1. O'reilly T (2007) What is web 2.0: Design patterns and business models for the next generation of software. Communications & strategies: 17.
2. Top A (2012) 500 global sites, 2011. Available: http://www alexa com/topsites.
3. Bodendorf F, Kaiser C (2009) Detecting opinion leaders and trends in online social networks. In: Proceedings of the 2nd ACM workshop on Social web search and mining. ACM, pp. 65–68.
4. Cammaerts B, Audenhove LV (2005) Online political debate, unbounded citizenship, and the problematic nature of a transnational public sphere. Political Communication 22: 179–196.
5. Zhang J, Ackerman MS, Adamic L (2007) Expertise networks in online communities: structure and algorithms. In: Proceedings of the 16th international conference on World Wide Web. ACM, pp. 221–230.

6. Abbasi A, Chen H (2005) Applying authorship analysis to extremist-group web forum messages. Intelligent Systems, IEEE 20: 67–75.
7. Benevenuto F, Rodrigues T, Cha M, Almeida V (2009) Characterizing user behavior in online social networks. In: Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference. ACM, pp. 49–62.
8. Yu J, Hu Y, Yu M, Di Z (2010) Analyzing netizens view and reply behaviors on the forum. Physica A: Statistical Mechanics and its Applications 389: 3267–3273.
9. Cattuto C, Loreto V, Pietronero L (2007) Semiotic dynamics and collaborative tagging. Proceedings of the National Academy of Sciences 104: 1461–1464.
10. Cooley RW (2000) Web usage mining: discovery and application of interesting patterns from web data. Ph.D. thesis, University of Minnesota.
11. Huberman BA (2009) Social attention in the age of the web. Working together or apart: Promoting the next generation of digital scholarship: 62.
12. Huberman BA, Pirolli PL, Pitkow JE, Lukose RM (1998) Strong regularities in world wide web surfing. Science 280: 95–97.
13. Johnson EJ, Bellman S, Lohse GL (2003) Cognitive lock-in and the power law of practice. Journal of Marketing: 62–75.
14. Bucklin RE, Sismeiro C (2003) A model of web site browsing behavior estimated on clickstream data. Journal of Marketing Research: 249–267.
15. Bucklin RE, Lattin JM, Ansari A, Gupta S, Bell D, et al. (2002) Choice and the internet: From clickstream to research stream. Marketing Letters 13: 245–258.
16. Bollen J, Van de Sompel H, Hagberg A, Bettencourt L, Chute R, et al. (2009) Clickstream data yields high-resolution maps of science. PLoS One 4: e4803.
17. Wu F, Huberman BA (2007) Novelty and collective attention. Proceedings of the National Academy of Sciences 104: 17599–17601.
18. Wu F, Wilkinson DM, Huberman BA (2009) Feedback loops of attention in peer production. In: Computational Science and Engineering, 2009. CSE'09. International Conference on. IEEE, volume 4, pp. 409–415.
19. Huberman BA, Romero DM, Wu F (2009) Crowdsourcing, attention and productivity. Journal of Information Science 35: 758–765.
20. Lerman K, Ghosh R (2010) Information contagion: An empirical study of the spread of news on digg and twitter social networks. ICWSM 10: 90–97.
21. West GB, Brown JH, Enquist BJ (1997) A general model for the origin of allometric scaling laws in biology. Science 276: 122–126.
22. Brown JH, Gillooly JF, Allen AP, Savage VM, West GB (2004) Toward a metabolic theory of ecology. Ecology 85: 1771–1789.
23. Zhang J, Guo L (2010) Scaling behaviors of weighted food webs as energy transportation networks. Journal of Theoretical Biology 264: 760–770.
24. Straškraba M, Jørgensen SE, Patten BC (1999) Ecosystems emerging: 2. dissipation. Ecological Modelling 117: 3–39.
25. Zhang J, Wu L (2013) Allometry and dissipation of ecological flow networks. arXiv preprint arXiv:13025803.
26. West GB, Brown JH, Enquist BJ (1999) The fourth dimension of life: fractal geometry and allometric scaling of organisms. Science 284: 1677–1679.
27. Banavar JR, Maritan A, Rinaldo A (1999) Size and form in efficient transportation networks. Nature 399: 130–132.
28. Garlaschelli D, Caldarelli G, Pietronero L (2003) Universal scaling relations in food webs. Nature 423: 165–168.
29. Cheng H, Cantú-Paz E (2010) Personalized click prediction in sponsored search. In: Proceedings of the third ACM international conference on Web search and data mining. ACM, pp. 351–360.
30. Fu X, Budzik J, Hammond KJ (2000) Mining navigation history for recommendation. In: Proceedings of the 5th international conference on Intelligent user interfaces. ACM, pp. 106–112.
31. Bejan A, Lorente S (2011) The constructal law and the evolution of design in nature. Physics of Life Reviews 8: 209–240.
32. Stephens MA (1974) Edf statistics for goodness of fit and some comparisons. Journal of the American statistical Association 69: 730–737.
33. Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world' networks. nature 393: 440–442.
34. Barabási AL, Albert R (1999) Emergence of scaling in random networks. science 286: 509–512.
35. Huberman BA (2003) The laws of the Web: Patterns in the ecology of information. MIT Press.
36. Bettencourt LM, Lobo J, Helbing D, Kühnert C, West GB (2007) Growth, innovation, scaling, and the pace of life in cities. Proceedings of the National Academy of Sciences 104: 7301–7306.
37. Leskovec J, Kleinberg J, Faloutsos C (2007) Graph evolution: Densification and shrinking diameters. ACM Transactions on Knowledge Discovery from Data (TKDD) 1: 2.
38. Cattuto C, Barrat A, Baldassarri A, Schehr G, Loreto V (2009) Collective dynamics of social annotation. Proceedings of the National Academy of Sciences 106: 10511–10515.
39. Tessone CJ, Geipel MM, Schweitzer F (2011) Sustainable growth in complex networks. EPL (Europhysics Letters) 96: 58005.
40. Wu L, Zhang J (2011) Accelerating growth and size-dependent distribution of human online activities. Physical Review E 84: 026113.
41. Henderson T, Bhatti S (2001) Modelling user behaviour in networked games. In: Proceedings of the ninth ACM international conference on Multimedia. ACM, pp.212–220.
42. Kim DH, Atluri V, Bieber M, Adam N, Yesha Y (2004) A clickstream-based collaborative filtering personalization model: towards a better performance. In: Proceedings of the 6th annual ACM international workshop on Web information and data management. ACM, pp.88–95.
43. Higashi M (1986) Extended input-output flow analysis of ecosystems. Ecological Modelling 32: 137–147.
44. Newman ME (2005) Power laws, pareto distributions and zipf's law. Contemporary physics 46: 323–351.