Scaling of human behavior during portal browsing

Anna Chmiel,[1] Kamila Kowalska,[2] and Janusz A. Hołyst[1]

[1]*Faculty of Physics, Center of Excellence for Complex Systems Research, Warsaw University of Technology, Koszykowa 75, PL-00-662 Warsaw, Poland*

[2]*Gemius S.A., Wołoska 7, PL-02-675 Warsaw, Poland*

We investigated flows of visitors migrating between different portal subpages. Two various portals were studied as weighted networks where nodes are portal subpages and edge weights are numbers of user transitions. Such networks differ from networks of portal subpages connected by hyperlinks prepared by portal designers. Distributions of link weights, node strengths, and times spent by visitors at one subpage follow power laws over several decades for data collected during two different days and for weekly data. The distribution of numbers $P(z)$ of unique subpages visited during one session is exponential and there is a square-root dependence between the total number of transitions n during a single visit and the average z. A model of portal surfing is developed where the browsing process corresponds to a self-attracting walk on the weighted network with a short memory. Results of numerical simulation are in agreement with weekly and daily portal data, and our analytical approach fits empirical data in the center part of scaling regime.

PACS number(s): 89.75.Da, 89.20.Hh, 89.75.Hc

I. INTRODUCTION

For physicists, the World Wide Web (WWW) is an intriguing complex object with hidden rules of dynamics that can be partially understood by observations of its specific statistics [1–15]. From the point of view of topology the WWW consists of the strong component, in and out components, and tendrils as well as disconnected clusters [1–4]. The problem of effective extracting of authoritative documents in WWW is crucial for web search engines and was considered already in 1999 by Kleinberg [16] from the point of view of graph theory. Several models of the WWW growth were developed based on concepts of complex evolving networks [1,2,5,6]. The examples are the preferential attachment process that leads to a scale-free distribution for degrees [5,7] observed in WWW documents or the model of redirected outgoing links [6] that well describes the supercritical distribution of clusters connected with the giant component. The WWW is, however, also a room of various *human actions* that can be easily investigated using the plethora of available data [15,17–20]. It was shown [15] that the origin of heavy tails of Internet traffic can be a natural feature of the optimal web design where an extended Shanon information theory is properly applied for portal architecture and file volume distributions. On the other hand Barabási [8] and Vázquez *et al.* [9] proposed that the bursty nature of various human activities observed in cyberspace (electronic mails and web browsing) is a consequence of decision-based queuing processes. Dezsö *et al.* [10] used this approach to explain the origin of distribution of time intervals between two consecutive visits of users at the Hungarian news portal. The extension of this model by a multitasks concept [11] gave a proper value for a characteristic exponent observed in portal data.

Another issue is the *navigability* of the whole web or its parts. Navigability is an attribute of web usability and it facilitates the way of finding specific information. It was recently discovered that optimal paths between two nodes in many complex networks can be found without the knowledge of a global network topology [21,22] due to the existence of hidden metric spaces. A lack of full information about the system is a common difficulty that is also present during the portal surfing [12,13]. Huberman *et al.* [12] described the browsing as a random selection of subpages that possess various attributes (values) for each user. Every next subpage attribute is stochastically related to the previous one. The resulting distribution of the number of clicks is the inverse Gauss function, which fits well to data collected from several American portals. The observation [12] was performed in 1997 when the portals were at the beginning of their development and users did not spend much time browsing them. In fact, the mean values of clicks per user observed in [12] were about 3. Today's portals are much larger and they are able to attract the visitor attention for a much longer time. Our data collected in 2007 and 2009 give a mean number of clicks of around 10 for studied portals.

The goal of this paper is to identify common strategies of portal browsing. The strategies are described by the way visitors navigate between subpages, how much time they spend on each of them, if and how many times they come back to the previously visited subpages, etc. Knowledge about these habits can also be a hint for portal designers to develop a marketing strategy that will fix the optimal number and distribution of advertisements. The usability of portal can be also dependent on the splitting of downloaded documents into several files [15]. In our study however we do not consider volumes of files downloaded by portal visitors since these volumes do not vary significantly at considered portals. We are also only loosely interested in the portal architecture defined by hyperlinks between various subpages. Of course visitors browsing habits are dependent on this architecture but as we shall show there is a significant fraction of visitors who jump between sites that are not directly hyperlinked.

The paper is organized as follows. In Sec. II, we present detailed details of our data set, and in Sec. III we discuss a portal structure with hyperlinks between distinct subpages. In Sec. IV the portal is treated as a set of pages that are

linked by visitor transitions and we analyze properties of the corresponding weighted network. Section V is focused on observed users' browsing strategies. A browsing model is introduced in Sec. VI as a special self-attracting walk, and complementary simulation results are discussed. An approximate analytical solution for our browsing model is presented in Sec. VII.

II. DATA SET

The analysis was based on cookie statistics provided by Gemius company for two Polish portals (names of the portals cannot be published because of data protection). A cookie is a small text file that every web browser automatically receives from a web server while visiting a webpage. It allows users to be differentiated and to maintain data related to them during navigation. The shortcoming of this method is that some users can delete their cookies and get a new one the next time they enter the Internet. Therefore, the number of cookie users does not correspond to the number of real Internet users. The difference is larger the longer the time of data collection, being negligible for 1 day, but immensely significant for 1 month. For both portals the data were collected in two time periods: on 27 July 2007 and two years later between 15 and 21 July 2009. In the last case the data were gathered for each day separately, as well as aggregated for the whole week.

The data collected for one website included user ID, subpage ID, time of user's arrival at a subpage (and, more precisely, the time she or he clicked on this page). It should be emphasized that the whole collection of times and subpages ID for a given user, which we will refer to as a "visit chain" in due course, cannot be identified with the time physically spent by the user on the subpages. This is a natural consequence of the fact that we detect users' activity only by their transition to new subpages and we do not have any information about their real behavior between two such events. In particular, we are unable to verify whether a user spent the time between two transitions reading the subpage content or whether he left it very quickly, started other activity (e.g., a phone call), and then came back to browse the Internet again.

III. PORTAL STRUCTURE

The structure of most webpages—and news portals in particular—is quite similar. The top of the structure and a gateway to the exploration of the whole portal is a main page. It is connected by hyperlinks with the most important subpages. These are usually topic services, related to sport, politics, culture, business, and so on. Every service is then divided into smaller subservices with more specific range of interest. For example, "sport" service can be divided into subservices dedicated to different sport disciplines that cannot be, however, reached from the main page. There are also bilateral hyperlinks between the subpages of the same level, directed links from the internal levels subpages to the main page and to some (but not all) subpages of the upper levels. In effect, the overall structure, although pretty complicated, can be divided into a few levels of interconnected internal
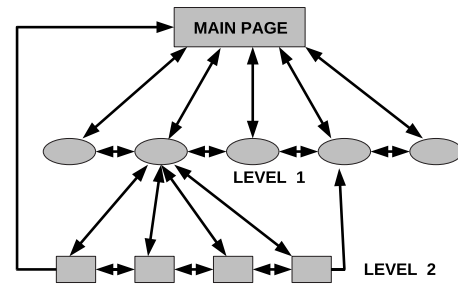

FIG. 1. Scheme of structure of hyperlink for portal.

subpages, with the main page at the top. We shall call the network of portal subpages connected by hyperlinks a *technical network* since this network is independent from the temporary user's behavior and it rather provides a user a possibility for easy jumps between subpages similarly as buses or trams give a possibility to travel between different stops in public transport networks [23]. Of course the design of the portal technical network should reflect needs of portal visitors similarly as the design of public transport networks should reflect needs of city inhabitants.

There are 458 subpages for portal A and 1371 for portal B (data from 2009) that can be monitored if a cookie user performs a click at one of them. However, considering the pages visited during 1 day, it turns out that only a fraction of this number is actually clicked (we will say "active") during that period. The exact number of active subpages varies slightly from day to day, being around 89% of all pages for portal A and 73% for portal B. The information about the subpages that have not been visited during a given period can provide some important insight for the webpage designers. A more precise analysis for portal A shows that about 17% of level 1 pages are not active, although in principle they are as easily reached from the main page as the actives ones. That observation would suggest that there are other important factors, beside the portal structure, that will make users visiting the certain subpages and omitting the others. The other result, coming from the comparison of the technical structure of portal A with a way the users browse on it, is that there is only a tiny fraction of user transitions from level 2 subpages to the main page. Although there exist hyperlinks allowing for such a navigating scenario (as presented in Fig. 1), they are hardly ever used in practice. There are also observed transitions of the users between the subpages that are not connected from the point of view of the technical structure of portal A. Such cases would correspond to observations of pedestrians walking between bus stops that are not directly connected. One of level 1 subpages has been chosen for such an analysis. It turns out that there are around 12% of no-hyperlink transitions from that subpage. 25% of them lead to some important subpages such as news, sports, and economy. There are several possible explanations of such phenomena: people use bookmarks to their favorite subpages, write the subpage URL by hand or use the "back" button for one-way connected subpages.

The above examples suggest that the very existence of a hyperlink between two subpages does not imply a link between these subpages, understood as a transition of a user
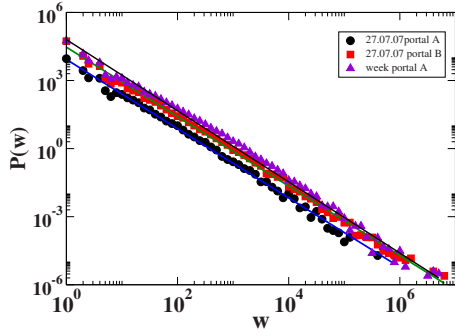
FIG. 2. (Color online) The weight distribution.

$$w_{ij} = m(i \rightarrow j) + m(j \rightarrow i). \tag{1}$$

One can observe a frequent habit to return to the subpage that was previously visited, so one visitor can pass many times over the same link. Therefore, the maximum link weight can be larger than the total number of users visiting the portal in the considered time period.

The weight distributions $P(w)$ (see Fig. 2) decay as a power law with the characteristic exponents η presented in Table I. Defining a node strength in the usual way,

$$s_i = \sum_j w_{ij}, \tag{2}$$

we found that the strength distribution is described by $P(s) \sim 1/s^\beta$ with β close to 1 (see Fig. 3). Both η and β exponents show *stability* for different 1 day intervals, as well as for 1 day and 1 week periods.

Our data set made us possible to measure the time spent by a user at one subpage if this time is understood as the time between two consecutive clicks on two different subpages performed by this user. Corresponding time distributions were analyzed by Dezsö *et al.* [10] and by Gonçalves and Ramasco [11], who found power-law relations with exponents $\gamma = 1.2$ and 1.25, respectively. The model of separately executed tasks [8] gives $\gamma = 1$ while the model of (bounded) tasks groups [11] leads to $\gamma > 1$. In our case, we measured

between them. Therefore, the network based on user activity can have different properties than the one based on the technical structure of the portal as it was also observed in [19].

IV. NETWORK STRUCTURE AND PORTAL TRAFFIC

We constructed a weighted network of subpages, defining a link weight as the number of users moving from one subpage (vertex) to another. We observed that the number of transitions from node i to node j, $m(i \rightarrow j)$, is roughly equal to the number of transitions in other directions $m(i \rightarrow j) \simeq m(j \rightarrow i)$; thus, we simplified the network topology by introducing undirected links with weights as follows:

TABLE I. Comparison of result obtained for different days. n_{user} is the number of users visiting portal during a considered period, N is the number of active subpages, η is the exponent of weighted distribution, w_{max} is the maximum weight, β is the exponent of strength distribution, s_{max} is the maximum of strength, n_{max} is the length of the longest visit, z_{max} is the maximum number of unique subpages visited by one user, and α is a exponent of distribution of numbers of unique subpages.

	n_{user}	N	η	w_{max}	β	s_{max}	n_{max}	z_{max}	α
Portal A									
27.07.07	1003673	195	1.52 ± 0.01	822433	1.02 ± 0.03	2767113	1729	41	0.41 ± 0.01
15.07.09	2272111	404	1.61 ± 0.01	1102903	1.05 ± 0.03	4072870	15800	76	0.30 ± 0.01
16.07.09	2270730	407	1.60 ± 0.01	1036696	1.05 ± 0.04	3937602	15806	74	0.33 ± 0.01
17.07.09	2149883	403	1.62 ± 0.01	828315	1.01 ± 0.04	3503820	14623	77	0.31 ± 0.01
18.07.09	1666424	400	1.63 ± 0.01	617187	1.07 ± 0.04	1948880	15711	44	0.31 ± 0.01
19.07.09	1940454	405	1.62 ± 0.01	687090	1.03 ± 0.04	2319760	15390	63	0.33 ± 0.01
20.07.09	2412961	413	1.60 ± 0.01	957636	1.06 ± 0.03	4005783	15809	79	0.30 ± 0.01
21.07.09	2331888	404	1.61 ± 0.01	1219132	1.06 ± 0.04	4216687	15682	82	0.29 ± 0.01
Weeks 15–21	8642106	413	1.57 ± 0.01	6160278	0.98 ± 0.03	25146528	108821	138	0.171 ± 0.005
Portal B									
27.07.07	3959624	515	1.53 ± 0.01	6857818	1.07 ± 0.03	25861104	7018	109	0.31 ± 0.01
15.07.09	5662539	1014	1.58 ± 0.01	4213204	1.13 ± 0.03	29056617	25081	121	0.29 ± 0.01
16.07.09	5546417	994	1.58 ± 0.01	4123942	1.12 ± 0.03	28154771	25270	113	0.28 ± 0.01
17.07.09	5285334	991	1.59 ± 0.01	3776969	1.12 ± 0.03	27336737	25194	124	0.27 ± 0.01
18.07.09	4026100	982	1.61 ± 0.01	3426243	1.12 ± 0.03	18776225	24407	76	0.28 ± 0.01
19.07.09	4493130	997	1.62 ± 0.01	3680387	1.11 ± 0.03	20863811	24166	105	0.30 ± 0.01
20.07.09	5664522	994	1.60 ± 0.01	4107782	1.11 ± 0.03	27572861	24368	95	0.30 ± 0.01
21.07.09	5693520	994	1.61 ± 0.01	3892358	1.11 ± 0.03	29085101	23785	112	0.28 ± 0.01
Weeks 15–21	17916433	1014	1.57 ± 0.01	27867827	1.02 ± 0.03	187079363	172247	222	0.142 ± 0.005
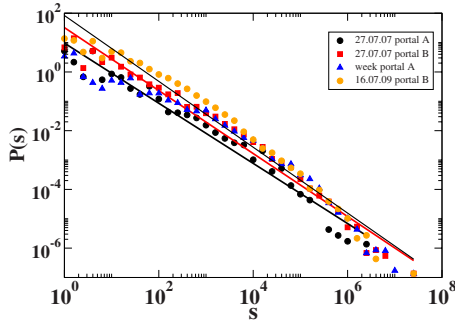
FIG. 3. (Color online) The strength distribution.

FIG. 5. (Color online) Distribution of numbers of unique subpages z visited by a user during one visit. The parameter α of exponential fitting is presented in Table I.

$\gamma = 1.27 \pm 0.01$ for portal A and $\gamma = 1.32 \pm 0.01$ for portal B (both results are for daily data), and the scaling was valid for the range over 2 decades (see Fig. 4).

To understand properties of weighted user transition network, we followed details of users' paths. We analyzed the distribution of numbers of unique subpages z visited by a user during a single visit (Fig. 5), and we observed an exponential behavior with a unique characteristic parameter α for a given portal,

$$P(z) = A \exp(-\alpha z). \tag{3}$$

The value of α parameter differs for different time intervals. The relation between α and the number of active subpages is straightforward and quite intuitive: the bigger the portal (i.e., the larger the number of the active subpages), the more unique subpages can be visited during one visit and, hence, the smaller is the value of α parameter.

Let us consider the relation between two variables: a number of jumps (transitions) between subpages n and the average number of distinct (unique) subpages $\langle z \rangle$ corresponding to the same fixed n value. One can refer to z as to the "interest horizon" since it describes the user's tendency to stick to a limited subset of certain subpages. The relation is presented in Fig. 6 and for $n \lesssim 100$ it can be described by the function

$$\langle z \rangle = a\sqrt{n}. \tag{4}$$

For both portals, we observed the same square-root dependence. Now, having Eq. (4) and the number distribution of

unique subpages, we can find the formula for the distribution of the number of jumps presented in Fig. 7. Since

$$P(n)dn = P(z)dz, \tag{5}$$

we get

$$P(n) = \frac{aA \exp(-\alpha a \sqrt{n})}{2\sqrt{n}}. \tag{6}$$

As we observed in Fig. 7 this formula fits to the collected daily data but for the weekly data the exponential behavior for the distribution of numbers of unique subpages appears after 20 steps (see Fig. 5). That is why the Eq. (6) fits only to the tail of the weekly data jump distribution (see Fig. 7).

V. VISITORS STRATEGY

The analysis of statistical properties of the visitor behavior reveals an important phenomenon that is frequent returns to a subpage previously visited. One can ask how a subpage at the portal affects the frequency of the return to this subpage. The most special subpage is the main page, in most cases being the starting point of browsing. Therefore, this subpage contains the biggest number of links. We assume that the main page is not the only subpage revisited during one visit chain. Let p^* be a return probability to a subpage visited one step earlier (meaning that a user reads the same subpage at the step numbers $n-1$ and $n+1$). We restricted
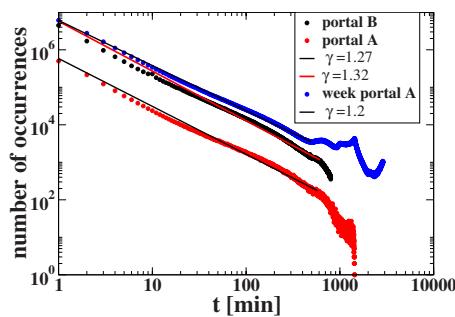

FIG. 4. (Color online) The distribution of time spent by the user on one subpage. For the weekly data the constraint of 2 days was imposed on the longest time interval between two consecutive clicks.
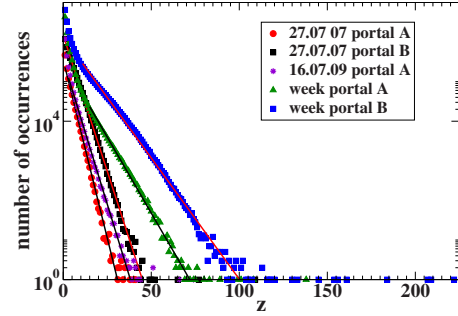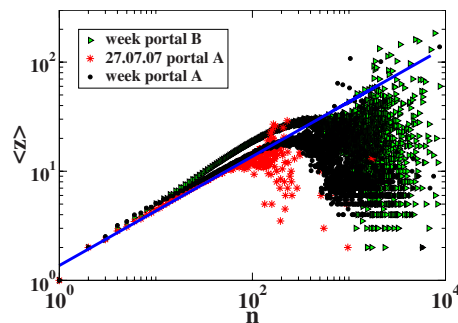

FIG. 6. (Color online) Relation between the average number of distinct subpages $\langle z \rangle$ and number of jumps n. Lines fitted with $\langle z \rangle = 1.44\sqrt{n}$ are the best for daily and weekly data from portal A; for weekly data from portal B the exponent increases to 0.52.
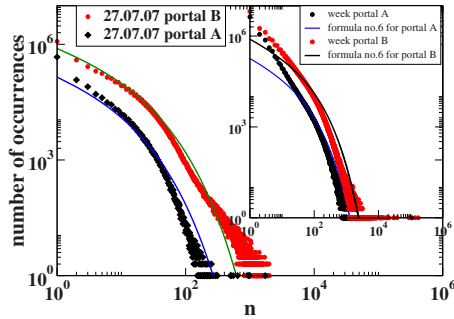
FIG. 7. (Color online) Jump distribution and fit to Eq. (6).

our calculation to the range of 100 steps only, where the scaling is observed. In Table II, we present the total return probability p^* and its two components that describe the coming back to the main page (p_{main}) and to a different subpage (p_{diff}). The total return probability is $p^* = p_{main} + p_{diff}$. The large value of p^* can be related to the use of the back button or opening of a new "window" and jumping between two open browser windows. For weekly data the value of p^* does not change a lot as compared to 1 day and it is equal to $p^* = 0.52$ for portal A and to $p^* = 0.57$ for portal B.

VI. SELF-ATTRACTING WALK AS A BROWSING SCENARIO

The relation between the average number of distinct subpages $\langle z \rangle$ and the number of jumps n can be interpreted as the relation between the average number of distinct visited sites and the number of steps (understood as time) in the problem of random walk. This classical problem of stochastic processes was considered by many authors (see, e.g., [24–29]). In one dimension, a random walk is characterized by the square-root relation between the number of distinct states and the number of visited states, $\langle z \rangle \sim \sqrt{n}$. From the famous Polya's theorem [27], it is known that the probability of returning (at any time) to the starting point by a random walker in a d-dimensional lattice can be less than 1 only for $d > 2$. In this sense, the dimension $d = 2$ is critical for this dynamics, $\langle z \rangle \sim n / \ln n$ [28]. For complex networks with scale-free degree distribution, there is $\langle z \rangle \sim n$ (see [26]).

Various models of *biased* random walks were analyzed (see, e.g., [30,31]). A special case is a model of self-attracting walk [32,33] where a state that was previously visited is preferred in the next time step but there are different scenarios of attracting relations. Here, we adopt such a model for portal browsing. The model reminds the process introduced by Fagin *et al.* [36] where the problem of random

TABLE II. Comparison of return probabilities data (27 July 2007) from portals A and B.

	Portal A	Portal B
Probability p^*	0.54	0.57
Probability p_{main}	0.29	0.27
Probability p_{diff}	0.25	0.30

walks with the back button was considered and it was shown that that such a process is an extension of Markovian chains. We took the topology of the network with weights between two nodes defined by Eq. (1). The dynamics on the network is very simple: each walker starts at the main page and then, with a transition probability p_{ij}, moves to one of the neighboring subpages. The probability of transition from vertex i to vertex j is proportional to the weight of this edge,

$$p_{ij} = \frac{w_{ij}}{\sum\limits_{k} w_{ik}} = \frac{w_{ij}}{s_i}. \tag{7}$$

After two initial steps a tendency of returning to a previously visited page is introduced as follows. At the step n, a walker returns to a node visited at the $n-2$ step with probability p^*, and with probability $1 - p^*$ he chooses a random neighbor, according to the transition probability p_{ij}. The results of such a simulation for both portals (Fig. 8) are close to the real data in the range of the first 30 steps for the 1 day data and the first 100 steps for the weekly data. One day simulation was based on the model parameters derived for the data from 15 July 2009 (404 active subpages). The real data presented in the figure corresponds to 21 July 2009 (also with 404 active subpages). The gray area in Fig. 8 shows the regime ($\langle z \rangle - \sigma_z(n), \langle z \rangle + \sigma_z(n)$), where $\sigma_z(n)$ is the standard deviation for distribution of number of unique visited subpages for a specific n. The standard deviation is increasing with n, except from the cases $\sigma_z(1) = \sigma_z(2) = 0$. In the whole regime the standard error is significantly larger than a difference between real data and numerical stimulation. The χ^2 test confirmed the agreement between empirical and numerical data in the range $2 < n < 300$ of statistical significance $\alpha = 0.01$. This results however from large values of standard deviations in empirical data. In Fig. 8 we also present the relation between z_{max} and the maximum number of unique subpages visited during the n-step visit. Over the first decade we observed a linear dependence until some critical point after which the maximum number of unique subpages is smaller, $5\sqrt{n}$. For the weekly data the critical point is around 22 for both portals.

The model of human behavior during the webpage browsing was presented by Gonçalves *et al.* [17]. This model applies to the behavior of users of quite specific portals, like university ones, and is based on the IP addresses. The data collected in [17] reflect the individual popularity of various subpages for various users. Such an analysis is possible to perform when the set of the users is relatively limited (for example, for the people from one university) and their IPs are known. In the case of the public portals the access to the IPs is not possible because of data protection. Then the analysis based on the cookies is more appropriate. There are some similar approaches in both models, such as assuming certain probability of returning to the visited pages or the possibility that a user can randomly select her or his next webpage. Our model of the self-attracting random walk with a short memory does not need the full information of history of special user, like in [17] where the bookmarks ranking are build; however, we use the global ranking of popularity subpage.
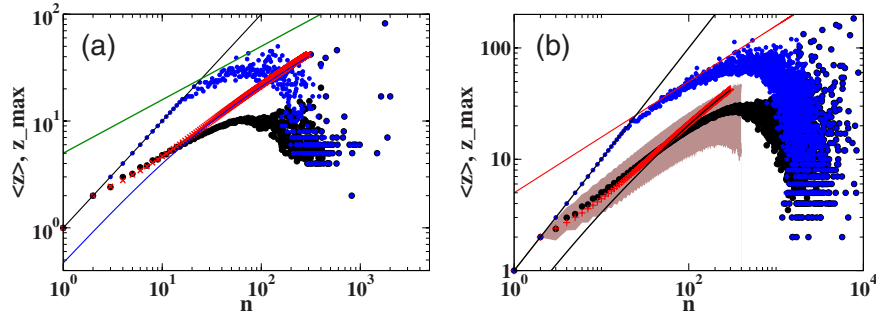
FIG. 8. (Color online) The relation between the average number of distinct subpages and the number of jumps. Left: 1 day data for portal A; right: weekly data for portal B. Black points represent data, blue points are the maximum number of unique subpages during n step visit, black lines fit to relation between the average number of distinct subpages and the number of jumps are Eq. (12); red crosses come from the numerical simulation. The gray area is the regime $(\langle z \rangle - \sigma_z(n), \langle z \rangle + \sigma_z(n))$. Results of numerical simulation (1×10^6 artificial users and n =300 steps for each) fit to the real data from portal A for number of steps $n < 30$, and for $n < 100$ for portal B. After system thermalization (more than 30 steps), we observed agreement between the analytical calculation and simulation results.

VII. ANALYTICAL APPROXIMATION

Let us now consider a random walk at the weighted network. In the infinite time limit, the stationary occupation probability ρ_i describing the probability that a walker is located at node i is given by [34,35]

$$\rho_i = \frac{s_i}{N\langle s \rangle}. \tag{8}$$

Unfortunately, in our case the walker lifetime at the network is not long enough to assume the stationary distribution of probability ρ_i because at the beginning of the walk the initial site is significant.

Despite this nonstationarity, we find an approximated relation between the average number of distinct subpages $\langle z \rangle$ and the number of jumps n. Let us define $d_s(n)$ as a fraction of vertices of strength s visited by a random walker at least once after n time steps. In our weighted networks, the relation between $\langle z \rangle$ and n is analogical to the unweighted networks discussed in [29],

$$\langle z \rangle(n) = N \sum_s P(s) d_s(n), \tag{9}$$

where N is the number of vertex subpages of the portal. Changes in $d_s(n)$ can be written as

$$\frac{\partial d_s}{\partial n} = [1 - d_s(n)]\rho(s). \tag{10}$$

Here, $\rho(s)$ is the probability that a walker observed in a random time moment is at site of strength s. The above equation is true when $\rho(s)$ is stationary, and it is only an approximation during first steps of the walker's path. Now we extend this equation by taking into account user's tendency to revisit subpages discussed in the previous section. Since the walker returns to the page visited two steps earlier with probability p^*, in such cases there is $\partial d_s / \partial n = 0$. It leads to the equation

$$\frac{\partial d_s^*}{\partial n} = [1 - d_s^*(n)]\rho(s)(1 - p^*). \tag{11}$$

Taking into account the initial condition $d_s^*(0) = 0$, we get the solution

$$d_s^*(n) = 1 - \exp[-n\rho(s)(1 - p^*)] \tag{12}$$

with the characteristic relaxation time

$$\tau(s) = \frac{1}{\rho(s)(1 - p^*)}. \tag{13}$$

One can see that the relaxation process is slowed down by the factor $1 - p^*$. From Eqs. (9) and (12), we have

$$\langle z \rangle(n) = N - N \sum_s P(s) \exp\left[-\frac{sn(1 - p^*)}{\langle s \rangle N}\right]. \tag{14}$$

Since for infinite networks there is a divergence of the first moment of the empirically observed distribution $P(s) \sim 1/s$, we used real data to estimate $\langle s \rangle = (1/N)\Sigma_{i=1}^N s_i$. The resulting solution (14) is presented in Fig. 8 and it fits with the numerical simulations discussed in Sec. VI.

VIII. CONCLUSIONS

We show that user's interest horizon, measured as the number of *distinct* visited subpages, is relatively small in comparison to the number of all transitions at the portals, i.e., to the number of all subpages visited by the user. This means that people return many times to the same subpage or pass by the same page during a 1 day and 1 week visit session. There can be various explanations for this phenomenon. The large probability of coming back to the main page can be a result of the technical portal structure (see Sec. III), with a main page being a hub of network. However, since the probability of returning to any other subpage is also significant, it can be suggested that it is somehow difficult for the users to find the information they need. So, if they consider a visited subpage inadequate to what they were expecting, they come back one step up the portal structure and try to go to

the other subpage. Since the number of distinct pages visited by portal users grows as a square root of the total number of clicks, the increase in newly visited pages is smaller and smaller for long time visits. It can be suggested that even the users that spend a lot of time at portal browsing are looking for a limited set of subjects. In this sense, the Internet seems to be a perfect tool to keep an eye on the changing situation in the regions of some importance to the users (for example, stock market indices, political news, and topical portals). Observation of week behavior of user argues for this explanation. However, the existence of such a global and easily accessed "knowledge mine" does not necessarily enlarge people's general interest horizon. The other possibility can result from the fact that portal visitors need to frequently pass over a few "transit" pages to come from one aim to another. Such a scenario would correspond well to the observed exponent $\gamma > 1$ (see Fig. 4) describing probability distributions of times between consecutive clicks. This scheme would also fit to the model of bounded group tasks proposed in [11]. Our simple model of a self-attracting walk shows that real data are in part reproduced by a short memory process. The observed scaling relation between an average number of distinct subpages $\langle z \rangle$ and a number of jumps n is approximately reconstructed using the strength of the node as a popularity range and the rule of coming back to the previous page with probability p^*. The solution of the rate equation fits with the simulation results for a number of clicks larger than $n > 30$. However, it is difficult to directly compare the developed model with the collected portal data because the number of users visiting more than 30 subpages is not large for daily data. For weekly data we observe that analytical results reflect the empirical data in the regime between 60 and 100. It suggests that a part of user activity can be modeled by random walkers with a short memory.

[1] R. Pastor-Satorras and A. Vespignani, *Evolution and Structure of the Internet: A Statistical Physics Approach* (Cambridge University Press, Cambridge, UK, 2004).

[2] S. N. Dorogovtsev and J. F. F. Mendes, *Evolution of Networks: From Biological Nets to the Internet and WWW* (Oxford University Press, New York, 2003).

[3] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener, Comput. Netw. **33**, 309 (2000).

[4] J. Kleinberg and S. Lawrence, Science **294**, 1849 (2001).

[5] P. L. Krapivsky and S. Redner, Comput. Netw. **39**, 261 (2002).

[6] B. Tadic, Physica A **293**, 273 (2000).

[7] A.-L. Barabási and R. Albert, Science **286**, 509 (1999).

[8] A.-L. Barabási, Nature (London) **435**, 207 (2005).

[9] A. Vázquez, J. G. Oliveira, Z. Dezsö, K. I. Goh, I. Kondor, and A.-L. Barabási, Phys. Rev. E **73**, 036127 (2006).

[10] Z. Dezsö, E. Almaas, A. Lukács, B. Rácz, I. Szakadát, and A.-L. Barabási, Phys. Rev. E **73**, 066132 (2006).

[11] B. Gonçalves and J. J. Ramasco, Phys. Rev. E **78**, 026123 (2008).

[12] B. A. Huberman, P. L. T. Pirolli, J. E. Pitkow, and R. M. Lukose, Science **280**, 95 (1998).

[13] Y. Zhou, H. Leung, and P. Winoto, IEEE Trans. Software Eng. **33**, 869 (2007).

[14] B. Kujawski, J. A. Holyst, and G. J. Rodgers, Phys. Rev. E **76**, 036103 (2007).

[15] X. Zhu, J. Yu, and J. Doyle, Proceedings of IEEINFOCOM, 2001 (unpublished), p. 1617.

[16] J. Kleinberg, J. ACM **46**, 604 (1999).

[17] B. Gonçalves, M. R. Meiss, J. J. Ramasco, A. Flammini, and F. Menczer, e-print arXiv:0901.38390.

[18] B. Gonçalves and J. Ramasco, e-print arXiv:0901.0498.

[19] M. Meiss, F. Menczer, S. Fortunato, A. Flammini, and A. Vespignani, Proceedings of WSDM, 2008 (unpublished), p. 6575.

[20] F. Radicchi, Phys. Rev. E **80**, 026118 (2009).

[21] M. Boguñá, D. Krioukov, and C. Claffy, Nat. Phys. **5**, 74 (2009).

[22] M. Boguñá and D. Krioukov, Phys. Rev. Lett. **102**, 058701 (2009).

[23] J. Sienkiewicz and J. A. Hołyst, Phys. Rev. E **72**, 046127 (2005).

[24] B. Huges, *Random Walks and Random Environments* (Clarendon Press, Oxford, UK, 1995).

[25] B. Huges and M. Sahimi, J. Stat. Phys. **29**, 781 (1982).

[26] J. D. Noh and H. Rieger, Phys. Rev. Lett. **92**, 118701 (2004).

[27] G. Polya, Math. Ann. **83**, 149 (1921).

[28] A. Dvoretzky and P. Erdös, in *Proceedings of the Second Berkeley Symposium on Mathematical Statistical and Probability*, edited by J. Neyman (University of California Press, Berkeley, 1950), p. 353.

[29] A. Baronchelli, M. Catanzaro, and R. Pastor-Satorras, Phys. Rev. E **78**, 011114 (2008).

[30] C. Anteneodo and W. A. M. Morgado, Phys. Rev. Lett. **99**, 180602 (2007).

[31] A. Fronczak and P. Fronczak, Phys. Rev. E **80**, 016107 (2009).

[32] H. E. Stanley, K. Kang, S. Redner, and R. L. Blumberg, Phys. Rev. Lett. **51**, 1223 (1983).

[33] A. Ordemann, G. Berkolaiko, S. Havlin, and A. Bunde, Phys. Rev. E **61**, R1005 (2000).

[34] S. M. Ross, *Stochastic Processes*, Wiley Series in Probability and Mathematical Statistics (Wiley, New York, 1996), pp. 203–213.

[35] T. M. Cover and J. A. Thomas, *Elements of Information Theory* (John Wiley and Sons., New York, 1991), pp. 66–69.

[36] R. Fagin, A. R. Karlin, J. Kleinberg, P. Raghavan, S. Rajagopalan, R. Rubinfeld, M. Sudan, and A. Tomkins, Ann. Appl. Probab. **11**, 810 (2001).