Networked User Engagement

Janette Lehmann
Universitat Pompeu Fabra
Barcelona, Spain
jnt.lehmann@gmail.com

Mounia Lalmas
Yahoo! Labs
Barcelona, Spain
mounia@acm.org

Ricardo Baeza-Yates
Yahoo! Labs
Barcelona, Spain
rbaeza@acm.org

Elad Yom-Tov
Microsoft Research
Herzliya, Israel
elad@ieee.org

ABSTRACT

Online providers frequently offer a variety of services, ranging from news to mail. These providers endeavour to keep users accessing and interacting with the sites offering these services, that is to *engage* users by spending time on these sites and returning regularly to them. The standard approach to evaluate engagement with a site is by measuring engagement metrics of each site separately. However, when assessing engagement within a network of sites, it is crucial to take into account the traffic between sites. This paper proposes a methodology for studying *networked used engagement*. We represent sites (nodes) and user traffic (edges) between them as a network, and apply complex network metrics to study networked user engagement. We demonstrate the value of these metrics when applied to 728 services offered by Yahoo! with a sample of 2M users and a total of 25M online sessions.

Categories and Subject Descriptors

H.1.2 [**User/Machine Systems**]: [Human information processing]

Keywords

User engagement; networks of sites; metrics; web traffic

1. INTRODUCTION

User engagement is the quality of the user experience that emphasises the positive aspects of the interaction, and in particular the phenomena associated with the desire to use a site [6]. Engagement is mostly measured through metrics aiming at assessing users' depth of interaction with a site. Widely-used metrics include click-through rates, time spent on a site, page views, return rates, number of unique users. This paper proposes a new set of metrics to study networked user engagement.

Many large online providers, such as AOL, Google, MSN and Yahoo!, provide a network of sites, offering, for instance, social media, news and mail services. Their aim is not only to engage users with each site, but across all sites in their network. We call user engagement with a network of sites *networked user engagement*. To achieve this, online providers spend increasing effort to direct users to various sites (e.g. using hyperlinks to help users navigate to and explore other sites); in other words, to increase *user traffic* between sites. In this context, although the success of a site still largely depends on itself, it also depends on how it is reached from other sites in the network. This leads to a virtuous cycle between site engagement and site traffic: each reinforces the other. In this paper, we focus on site traffic.

User traffic has been studied in a number of web contexts [1, 4], for example looking at the traffic between sites, the return rate, how users visit web pages, and the effect of these on, for instance, site popularity. The effect of the traffic on user engagement with respect to a network of sites has however not been studied. When assessing the engagement of a site, accounting for user traffic is not new. For instance, analytic companies such as `alexa.com` produce statistics about the incoming and outgoing traffic of a site. However, the focus is the traffic to and from a site, and not the traffic within a network of sites and its effect on user engagement. In addition, it remains unknown whether and how the traffic between sites, the *network effect*, influences engagement within a network of sites.

How can we measure networked user engagement? Current engagement metrics do not account for the user traffic in a provider network. They were designed to measure engagement at site level, and how to apply or adapt them to measure networked user engagement is not obvious. We present a methodology to measure networked user engagement. For that, we model sites (the nodes) and user traffic between them (the edges) as a network, and use metrics from complex network analysis [5] to enhance the understanding of user engagement within a provider network. In this paper, we focus on the measurement of user traffic, and the insights gained in doing so.

2. DATA AND NETWORKS

We collected one month (July 2011) of interaction data (triples of browser cookie, URL, and timestamp) from a sample of 2M users who gave their consent to provide browsing data through the toolbar of Yahoo!. A total of 728 sites were selected encompassing diverse types of services such as news, weather, e-commerce, email, and movies. We refer to these as the *provider network* (in our case Yahoo! network); all other sites were considered as external. The user activities

UEO'13, November 1, 2013, San Francisco, CA, USA.
Copyright 2013 ACM 978-1-4503-2421-2/13/11
http://dx.doi.org/10.1145/2512875.2512877 ...$15.00.

were split into sessions, where a session ends if more than 30 minutes have elapsed between two successive activities. We identified around 25M sessions.

Traffic networks Our aim is to provide insights into user engagement with respect to the Yahoo! network of sites. We therefore used the selected 728 sites from that company to extract a number of weighted directed networks $G = (N, E)$, where the set of nodes N corresponds to sites and the set of edges E corresponds to the user traffic between them. The edge weight $\lambda(e)$ between site (node) n_A and site (node) n_B represents the size of the user traffic between these two sites, which we define as the number of clicks from n_A to n_B. To remove unimportant connections (low traffic), we exclude all edges for which there were less than a fixed threshold, in our case, of 10 clicks between the two sites. Although the choice of thresholds will affect the network structure, we want to focus on edges for which we observe enough traffic, and thus can have an effect on networked user engagement. Finally, as many sessions involve sites outside the provider network,[1] we defined two types of edges: e_{int} corresponds to user navigating between two Yahoo! sites (internal traffic); and e_{ext}, corresponds to user traffic from a Yahoo! site to external sites, but returning to another Yahoo! site within the same session (return traffic). We exclude traffic returning to the same site, as our focus is on the traffic between sites.

Network instances We define 12 weighted directed networks to study the different factors that may affect networked user engagement. The networks are listed in Table 1.

The *edge-based* network $Edge_{int}$ is based on all Yahoo! sites and the internal traffic (e_i) for the month of July 2011. All other networks are variations of this network. The network $Edge_{int+ext}$ considers both internal and returning traffic (e_{ext}), which allows us to study the effect of leaving the provider network. Previous work [2] has shown the importance of user loyalty and temporal aspects when studying user engagement. We therefore defined three *user-based* networks according to a loyalty criteria, measured by the number of active days within a month per user. We refer to them as Causal, Normal and VIP networks. We also define two *time-based* networks based on the traffic for two separate days in our data period. We decided to compare the traffic created on a Wednesday ($Time_{wed}$) with the traffic created on a Sunday ($Time_{sun}$), i.e. a week day compared to a weekend.

Finally, five *country-based* networks were created, where each contains the sites and traffic for one country. The five countries are located in Europe to avoid "major" differences in user browsing behaviour based on cultural differences. We selected the 22 most popular sites that have a counterpart in most of the five countries as the set of sites for each of the networks. Looking at specific countries investigates whether networked user engagement varies across countries.

3. NETWORK EFFECT

Existing research demonstrated that, for instance, a page popularity depends on the user traffic between pages [3]. The same can be observed with respect to site popularity. We show that an effect can be observed for the sites forming a network (network effect), where the number of nodes is significantly smaller, but the number of edges much larger than in typical page-level browsing networks. We also show the reverse effect; site engagement has an effect on the traffic (site effect).

Table 1: Networks under consideration.

$Edge_{int}$ $Edge_{int+ext}$	Edge type e_{int} Edge type e_{int} and e_{ext}
$Time_{wed}$, $Time_{sun}$	Two days in July
$User_{causal}$ $User_{normal}$ $User_{VIP}$	1-4 active days 5-15 active days >15 active days
$Country_1$, ... , $Country_5$	5 countries from Europe

For the *edge-based* network $Edge_{int}$, we calculate the popularity of each site as the number of users that visited that site. The network traffic is made of incoming and outgoing traffic for each node (site) in the network. We use linear regression to measure the correlation between incoming traffic and the popularity of a site (we call this the *network effect*) and the popularity of a site and the outgoing traffic (we call this the site effect). The correlation coefficients are $R^2 = 0.85$ (network effect) and $R^2 = 0.84$ (site effect) with $p < e^{-5}$, revealing that the effect is quite significant in both cases: site popularity is highly correlated with incoming traffic, and vice versa. As both are highly correlated, and the differences between them are not statistically significant (using a ranksum test), we cannot conclude whether the site effect or the network effect is higher. However, we can conclude that there is a *network effect*. Looking at the other networks, the various factors considered (time, country, user loyalty), did not have any influence on the network nor the site effect.

4. NETWORK METRICS

We use standard complex network metrics to gain information about the structure of our 12 networks. The metrics *density* and *reciprocity* provide information about the network connectivity, whereas *subNW* and *GCC* reflect the network modularity.

TrafficVolume This is the sum of the edge weights, where the weight is the number of clicks between two sites. A high value is a sign of a high networked user engagement; users navigate often between the sites of the network.

Density This is the ratio between the number of edges compared to the number of all possible edges, and describes the connectivity of the network. A high connectivity indicates a high networked user engagement; users navigate between many different sites.

Reciprocity This is the proportion of edges for which the traffic is going in both directions. A high reciprocity can be interpreted as a high networked user engagement; users do not only navigate from one site to another, they also tend to return to previously visited sites.

Subnetworks [$subNW_{mod}$] A network with a high modularity contains densely connected subnetwork (modules) that are sparsely connected with each other. We use a random walk approach to calculate modularity. A high modularity indicates that users visit many sites of one subnetwork, but hardly navigate to other subnetworks. This could be observed in the networks that contain sites from all countries; users navigate most of the times between the sites forming

[1] Within a session, 1.7 Yahoo! sites and 4.9 external sites were accessed on average.

Table 2: Characterising the user traffic for the 12 networks. The values for the edge-, time- and user-based networks are combined. The average and standard deviation are shown $[avg|sd]$.

<table>
<tr><th></th><th>Edge, Time, User</th><th>Country</th></tr>
<tr><td>TrafficVolume</td><td>131M | 144M</td><td>3.7M | 2.5M</td></tr>
<tr><td>Density</td><td>0.017 | 0.011</td><td>0.303 | 0.029</td></tr>
<tr><td>Reciprocity</td><td>0.826 | 0.022</td><td>0.913 | 0.014</td></tr>
<tr><td>SubNW</td><td>0.465 | 0.057</td><td>0.006 | 0.057</td></tr>
<tr><td>GCC</td><td>0.546 | 0.060</td><td>0.754 | 0.048</td></tr>
</table>

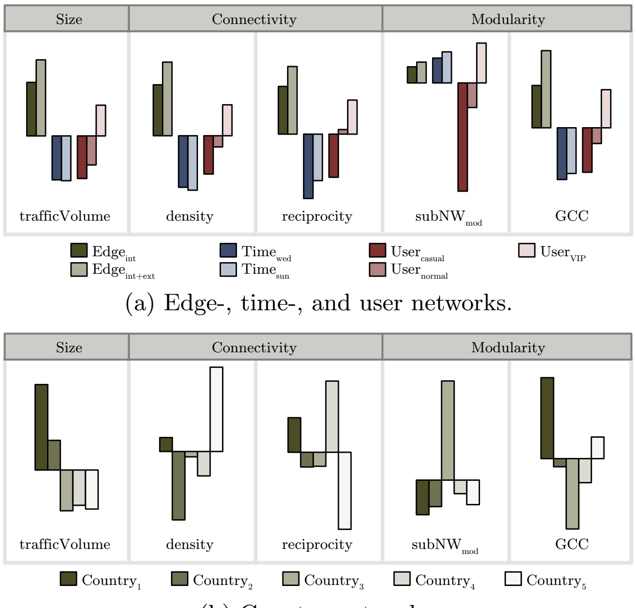
one country, a country subnetwork. A low modularity indicates that the connectivity between the sites of a subnetwork is low, a sign of low networked user engagement. However, a low modularity can also suggest that users navigate between all sites in the network, i.e. all nodes in the network are well connected. This could be the case for any of the five country networks studied in this paper. The *density* allows to distinguish between the two cases of low modularity.

Global Clustering Coefficient $[GCC]$ This metric captures the existence of tightly connected triples of sites. We use a generalised version of the GCC, which accounts for the edge weight (number of clicks) between the sites. For instance, the front pages[2] in Yahoo! are sites that play an important role in directing traffic to other sites. A high GCC would indicate that front pages are less used to navigate between sites. Instead, the sites are accessed directly, which could be considered as a higher networked user engagement; users are “aware” of these sites and do not need to be directed to them through, for instance, front pages.

5. MEASURING NETWORKED USER ENGAGEMENT

We investigate the extent to which the 12 constructed networks differ in terms of their user traffic. The country-based networks are analysed separately, since they contain only a subset of sites, 22 sites, from the original network.

Network characteristics We study the characteristics of our networks using the five metrics under consideration. The average and standard deviation of each metric is shown in Table 2. $TrafficVolume$ is multiplied by a constant factor for confidential reasons. We observe less traffic for country-based networks ($TrafficVolume$ is $3.7M$ compared to $131M$) as we would expect (these are smaller networks). The standard deviation for the edge-, time- and user-based networks is very high ($sd = 144M$), as, for instance, the time-based networks are based on the traffic of one day, whereas the edge-based networks are created with the traffic of a whole month. The density for country-based networks is higher (*density* is 0.303 compared to 0.017). Edge-, time- and user-based networks contain nodes (sites) of many different countries and since users usually navigate between the sites of one country, only few edges between sites of different countries exist. This explains also the higher modularity of the edge-, time and user-based networks ($SubNW$, which is 0.465 compared to 0.006). The networks contain densely connected subnetworks, one for each country, whereas for the country-based networks, users usually navigate between most sites. Because of this, the Global clustering coefficient (GCC) and *reciprocity*, although high in general, are also higher for country-based networks.



(a) Edge-, time-, and user networks.

(b) Country-networks.

Figure 1: Differences in networked user engagement.

Next, we compare the various networks to study networked user engagement. We computed the z-score for each metric. The corresponding values are plotted in Figure 1.

Edge-based networks We compare $Edge_{int}$, which includes the internal traffic only, and $Edge_{int+ext}$, which considers the traffic between the provider sites when navigating over external sites (return traffic).

We observe a significant increase in the traffic volume ($trafficVolume$) when accounting for returning traffic: the total traffic increases by 18.26%. If we consider loop edges as well, i.e. cases in which a user leaves the provider network, but returns to the same site in the network, an increase of 37.12% can be observed. The additional traffic increases *density* and *reciprocity*. This suggests that users leave the network from some site A and return to some other site B, but never navigate directly from site A to B within the network. The modularity of the return traffic network increases ($subNW_{mod}$ and GCC), implying that well connected parts of the network become even denser when accounting for returning traffic. This suggests that, when users leave the provider network, but return within the same session, they tend to stay within the same part of the network.

We can infer that leaving the provider network does not necessarily entail less networked user engagement. Also, not only users often return to the network, but many sites become connected through returning traffic. The fact that edges connect two sites only when considering the returning traffic could point to missing hyperlinks in the provider network. Adding these hyperlinks could increase networked user engagement, as they may help users navigating between sites, and thus staying longer in the network.

[2] The 728 services are from 80 countries, each of them with a front page to direct users to the various services offered to that country.

Time-based networks $Time_{wed}$ and $Time_{sun}$ are two time-based networks representing the traffic on a Wednesday and a Sunday, respectively. Although there are less connections (*density*) and traffic in the network on Sunday, the *reciprocity* and *modularity* ($subNW_{mod}$ and GCC) are higher for that network. This suggests a more leisurely browsing behaviour on the weekend. Users seem to be browsing with less specific goals in mind and therefore return more often to already visited sites and visit more sites than usual. Users tend to stay in one part of the network, which is likely to reflect that they visit sites that belong to one country. During the week (in our case a Wednesday), user navigations seem to be more goal-oriented; users visit the provider network to check their mails or search for specific information.

User-based networks We compare how networked user engagement changes with respect to the loyalty of a user. VIP users navigate more frequently between sites (higher $trafficVolume$) than those from other user groups. Also, it seems more probable that VIP users navigate in both directions between two sites (higher *reciprocity*), whereas, for instance, causal users go from one site to another but returned less to the previous site. We speculate that causal users are 'aware' of and/or are interested in a smaller set of sites (lower *density*), which they accessed mainly from front pages (low GCC), whereas VIP users are not only more engaged within one site, but with several sites (high $subNW_{mod}$).

Country-based networks The objective here is to show that the networks from the different countries vary in their networked user engagement (user traffic). Using network metrics we are able to compare engagement between provider networks, as we do it when comparing sites using engagement metrics. The differences in Figure 1(b) demonstrate that network metrics provide different insights into how users engage with the whole provider network.

The highest networked user engagement can be observed for the first country ($County_1$). This country has the highest traffic volume, the highest GCC, and the second highest *density* and *reciprocity*. This shows that users navigate often between many sites in that country network and do so in both directions; a sign of high networked user engagement. The modularity $subNW_{mod}$ is the lowest, indicating that there are no subsets of sites that are used independently from the other sites.

We look now at the network structure of the last country ($County_5$). Although the traffic volume between sites of this country is low (low $trafficVolume$), there is traffic between many of the sites (highest *density*, high GCC). We can also observe that users do not tend to navigate between sites in both directions (lowest *reciprocity*). However, the fact that the modularity metric $subNW_{mod}$ is low suggests that many sites of that country network are visited. Considering that we could observe a network and site effects (Section 3) we assume that this country (we verified this was in fact true) has a lower number of users compared to the other four country networks. However these users are engaged with many sites, implying a high networked user engagement.

The lowest networked user engagement can be observed for the third country ($County_3$). Although users tend to navigate in both directions between sites (average *reciprocity*) and an average number of edges exists (average *density*), we observe a low traffic volume and the modularity $subNW_{mod}$ is high, implying that there are several network parts (sets of sites) with their own users. We speculate that the users are not "aware" of the other sites and that they rely on links, for instance from the front page site of the country, to navigate to other sites (low GCC) of that country network.

6. CONCLUSIONS AND FUTURE WORK

This paper proposed a methodology for studying networked used engagement, that is, user engagement of a network of sites, such as those operated by large Internet companies such as AOL, Google, MSN and Yahoo!. Networked user engagement considers user traffic between services and we use complex network metrics to study it. We demonstrated the value of these metrics when applied to 728 services offered by Yahoo! with a sample of 2M users and a total of 25M online sessions. We observed that, for instance, VIP users access more sites and navigate more between them, leaving a network of sites does not mean less engagement, and finally user traffic varies across countries.

We did not apply complex networks metrics that calculate the centrality of nodes. These metrics can bring further insights in networked user engagement, and in particular when compared and combined with engagement metrics (such as those listed in the introduction). We also did not differentiate between the ways users navigate between services, whether clicking on a link, using the back button, or using a bookmark. We can build additional types of networks where, for instance, two sites are connected if there are web links between them. The same methodology can bring insights into how web links influence networked user engagement. Finally, the source of the traffic (e.g. users that come from a news or social media site) and its effect on networked user engagement is something that is important to investigate.

Our ultimate aim is to understand how a network of sites and the traffic between sites can be optimised to "better" engage users. The methodology presented in this paper constitutes the first step towards achieving this aim.

7. ACKNOWLEDGMENTS

This work was partially funded by Grant TIN2009-14560-C03-01 of the Ministry of Science and Innovation of Spain.

8. REFERENCES

[1] A. Chmiel, K. Kowalska, and J. A. Hołyst. Scaling of human behavior during portal browsing. In *Phys. Rev. E*. 2009.

[2] J. Lehmann, M. Lalmas, E. Yom-Tov, and G. Dupret. Models of user engagement. In *Proc. UMAP*, 2010.

[3] Y. Liu, B. Gao, T.-Y. Liu, Y. Zhang, Z. Ma, S. He, and H. Li. Browserank: Letting web users vote for page importance. In *Proc. SIGIR*, 2008.

[4] M. Meiss, B. Goncalves, J. Ramasco, A. Flammini, and F. Menczer. Modeling traffic on the web graph. In *Proc. WAW*. 2010.

[5] M. E. J. Newman. The structure and function of complex networks. In *SIAM Review*, 2003.

[6] H. L. O'Brien and E. G. Toms. What is user engagement? A conceptual framework for defining user engagement with technology. *ASIS*, 2008.