Measuring Inter-Site Engagement in a Network of Sites

Janette Lehmann[a], Mounia Lalmas[b], Ricardo Baeza-Yates[c]

[a] *Universitat Pompeu Fabra, Barcelona, Spain; lehmannj@acm.org*
[b] *Yahoo Labs, London, UK; mounia@acm.org*
[c] *Yahoo Labs, Sunnyvale, USA; rbaeza@acm.org*

Abstract

User engagement is a key concept in the design of websites, motivated by the observation that successful websites are not just used, but are engaged with. Engagement metrics enable us to perform large-scale web analytic studies to understand how users engage with a website. Many large online providers (e.g., AOL, Google, Yahoo) offer a variety of websites, ranging from shopping to news. Standard engagement metrics are not able to assess engagement with more than one website, as they do not account for the user traffic between websites. We therefore propose a methodology for studying inter-site engagement by modelling websites (nodes) and user traffic (edges) between them as a network. Our methodology reduces the complexity of the data, and hence metrics can be efficiently employed to study user engagement within such networks. The value of our approach was demonstrated on 228 websites offered by Yahoo and a sample of 661M online sessions.

Keywords: user engagement, network of sites, metrics, inter-site engagement, large-scale web analytics

1. Introduction

User engagement refers to the quality of the user experience associated with the desire to use a website (O'Brien and Toms, 2008). In the online industry, engagement is measured through online behaviour metrics aiming at assessing users' depth of interaction with a site. Widely used metrics include number of visits, click-through rates, time spent, and page views on a site. Since these metrics are able to process large volumes of data in real-time, they are extensively used by the web analytics community and Internet market research companies such as comScore as proxy for online user engagement. In this chapter, we also study user engagement using online behaviour metrics, which we refer to as *engagement metrics*, but with respect to a network of sites.

Many large online providers (e.g., AOL, Google, Yahoo) offer a variety of sites, ranging from shopping to news. The aim of these large online providers is not only to engage users with each site, but with as many sites as possible. These providers spend increasing effort to direct users to various sites, for example by using hyperlinks to help users navigate to and explore other sites in their

network; in other words, they want to increase the users traffic between sites. In this context, although the success of a site still largely depends on itself, it also depends on how and whether it is reached from other sites in the network. This leads to a strong relationship between site engagement and site traffic: each reinforces the other. We refer to this as the *network effect*.

When assessing the engagement of a site, accounting for user traffic is not new. For instance, search engines are major sources of referrals, as are social media sites. Knowing how users arrive at a site is used to optimise the site, e.g. by choosing better keywords (search engine optimisation). Web 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 between sites and its effect on user engagement.

In addition, engagement metrics cannot measure such online behaviour, as they do not account for the traffic (interactions) between sites. They were designed to measure engagement at site level, and how to apply or adapt them to measure engagement in a provider network is not obvious. We therefore propose a methodology for studying *inter-site engagement*; user engagement within a network of sites. We model sites (nodes) and user traffic (edges) between them as a network, and employ inter-site metrics to measure the engagement with respect to a whole provider network of sites. Since it is unknown whether and how the traffic between sites influences user engagement, we also employ inter-site metrics at node (site) level to study the relationship between site and inter-site engagement. Some of the metrics are borrowed from the area of complex network analysis (Newman, 2003), for instance, density at network-level, and page rank at node-level. We then perform a large-scale study that shows how these metrics can be used and how they enhance our understanding of user engagement within a network of sites.

Our approach allows us to study user engagement at a large-scale while accounting for the relationships between sites. First, we reduce the complexity of the browsing data by transforming it into a network. Second, we can efficiently calculate inter-site metrics using a single processor machine, because the resulting network is usually small. In case of a large network, approaches performing large-scale complex network analysis can be used (e.g. Xue et al., 2010).

We start by covering previous work in Section 2. The inter-site metrics used in our work are introduced and evaluated in Sections 3 and 4, respectively. Section 5 analyses how returning traffic (users leaving the network but returning within the same session), the loyalty of users, and other aspects affect inter-site engagement. The network effect is investigated in Section 6. Section 7 analyses how hyperlinks influence inter-site engagement. The chapter ends with conclusions and thoughts for future work.

To the best of our knowledge, this is the first study looking at user engagement from a network perspective, and at possibilities to measure it.

2. Related work

Our work spans across several research areas. We discuss them, and position our work in their context.

Web analytics and user engagement. User engagement is the quality of the user experience associated with being captivated by an application, and so being motivated to use it (O'Brien and Toms, 2008). Users do not just use an application, they engage with it. The web analytics community and the online industry have studied user engagement through online behaviour metrics that assess users' depth of engagement with a site. Several reports (e.g. Peterson and Carrabis, 2008) contain studies on existing online behaviour metrics and their usage. Widely used metrics include click-through rates, number of page views, time spent on a site, how often users return to a site and number of users per month. Only online behavioural metrics are scalable to millions of users, and are commonly employed as a *proxy* for user engagement to websites. Two million users accessing a site daily is a strong indication of a high engagement with that site. Our work adds to online behaviour metrics, which we refer to as engagement metrics, metrics accounting for user behaviour within a provider network of sites.

Browsing behaviour. User traffic on the Web has been studied in a number of contexts, for example looking at the general browsing characteristics of users (Kumar and Tomkins, 2010; Meiss et al., 2010; Cockburn and McKenzie, 2001), how users visit websites (Bucklin and Sismeiro, 2003), the return rate to a website (Lehmann et al., 2013; Dupret and Lalmas, 2013), or how users discover and explore new sites (Beauvisage, 2009). From these and other studies, several user navigation models were developed (Meiss et al., 2010; Simkin and Roychowdhury, 2008; Chmiel et al., 2009), for example accounting for the usage of bookmarks, back buttons, teleportation, etc. These models, based on formalisms such as branching processes, aimed to understand how users access sites and pages within them, and its effect on, for instance, link traffic and site popularity (Meiss et al., 2010; Simkin and Roychowdhury, 2008), and loyalty to a site (Cockburn and McKenzie, 2001; Beauvisage, 2009).

Several studies focused on the browsing behaviour across sites and how to support such browsing behaviour (Koidl et al., 2014). Jiang et al. (2012), Johnson et al. (2004) and Park and Fader (2004) analysed cross-site navigation in the context of online shopping, and showed, for instance, that online shoppers do some research on products they wish to purchase (e.g., on social media sites) before actually purchasing them and that the more active shoppers tend to visit more shopping sites. The PEW Research Center (2010, 2012) studied the reading behaviour across news sites. The studies found that 57% of users routinely obtain their news from between two to five news sites. Moreover, social media and search sites are playing an important role in the consumption of news. Dellarocas et al. (2013), De Maeyer (2011) and Roos (2012) even advocate that hyperlinks to other news sites can increase profits in a costless way, because doing so provides a more interactive, credible, transparent, and diverse news reading experience.

The main focus of our work is the development of metrics that are able to capture various aspects of cross-site navigation behaviour, but between sites belonging to the same provider.

Network analysis. In this work, we propose to incorporate user traffic into the study of user engagement by modelling sites and the traffic between them as a network. We use metrics from the area of complex network analysis (Newman, 2003) together with engagement metrics to study *inter-site engagement*. Additionally, we study the impact of the hyperlink structure in the provider network on inter-site engagement. Many types of complex networks have been studied (Newman, 2003); those closer to our work are user traffic networks (Chmiel et al., 2009; Meiss et al., 2008; Wu et al., 2011) investigating the traffic between web elements (pages, hosts, and sites), and hyperlink networks (Barabási et al., 2000) studying the topological (link) structure of the Web. Research on user traffic networks has looked at the structure of the networks (using metrics such as degree distribution) and evolution (Chmiel et al., 2009; Meiss et al., 2008). Some studies also considered engagement metrics (Chmiel et al., 2009).

There are many ranking algorithms that incorporate the browsing behaviour within a network or its hyperlink structure to rank pages (Page et al., 1999; Liu et al., 2008), images (Trevisiol et al., 2012), news articles (Trevisiol et al., 2014), etc. A famous ranking algorithm is PageRank (Page et al., 1999), which ranks pages using the structure of the hyperlink network. However, Liu et al. (2008) showed that incorporating the time that users spend on a page is outperforming the solely link-based approaches. We extend existing research by developing a measure that ranks sites in a provider network according to how much users engage to the network after visiting that site. This can be compared to the problem of influence maximisation (Kempe et al., 2003; Chen et al., 2009; Cheng et al., 2013), i.e., the identification of the n most influential nodes in, for instance, social or epidemic networks. In our context, we identify nodes (sites) that maximise the engagement to the network of sites. The measure itself is an adaption of the downstream metric of Yom-Tov et al. (2013) to our network model.

3. Data, networks and metrics

Our study is based on anonymised browsing data (tuples of browser cookie, URL, referring URL, and timestamp) collected over a period of 12 months (August 2013 to July 2014) from a sample of users who gave their consent to provide browsing data through the toolbar of Yahoo. The user activities were split into sessions, where a session ends if more than 30 minutes have elapsed between two successive activities, a common way to identify the end of a session (Catledge and Pitkow, 1995). We extracted a sample of 661M sessions. Each session consists of page views on Yahoo and other sites (e.g., facebook.com, cnn.com). The browsing activity of a user on Yahoo during a session was used to create several provider (in our case Yahoo) networks as described in the following subsection.

To handle the large volume of browsing data, we used the Hadoop Map/Reduce

Table 1: Network instances based on five country in one continent and the US network. For each network, we provide the number of network instances, and the average and standard deviation of the number of clicks per instance.

Network	Number of instances	Clicks per network
Monthly-based networks (February 2014).		
INT_{Feb}	1	16M
US_{Feb}	1	235M
Daily-based networks (August 2013 - July 2014).		
$INT_{01/08}$, ..., $INT_{31/07}$	356	577K|230K
$US_{01/08}$, ..., $US_{31/07}$	356	8.6M|2.9M
Country-based networks.		
INT_{c1}, ..., INT_{c5}	5	3.2M|3.6M

framework[1] to pre-process the data and extract the provider networks. After pre-processing the data, each reducer job created a sub-network based on a chunk of browsing data. Finally, the sub-networks were merged into one network. Since the resulting provider networks were small, we then used R[2] in conjunction to the igraph[3] and tnet[4] packages to calculate the metrics for each of the networks.

3.1. Provider networks

Our aim is to provide insights into user engagement with respect to Yahoo's network of sites. We created two provider networks using a total of 228 sites encompassing diverse services such as news, mail, and search.[5] The first provider network consists of a selection of 73 sites based in the Unites States, whereas the second network is based on sites from five countries from the same continent. We selected the 31 most popular sites that have a counterpart in all of the countries, resulting in 155 sites in total. The provider networks are weighted directed networks $G = (N, E)$, where the set of nodes N corresponds to sites and the set of edges E to the user traffic between them. The edge weight $w_{i,j}$ between node (site) n_i and node (site) n_j represents the size of the user traffic between these two nodes, which we define as the number of clicks from n_i to n_j. Whereas the nodes in the network are fixed, the edge weights depend on the selected browsing data. This allows us to create different network instances.

Network instances. The network instances are listed in Table 1. We defined various network instances of the INT network to evaluate the metrics described at the end of this section. The metrics characterise the inter-site engagement between nodes or within the whole network. The metrics are evaluated in Section 4. We extracted browsing data from each day between August 2013 and

[1] `http://hadoop.apache.org/`

[2] `http://www.r-project.org/`

[3] `http://igraph.org/r/`

[4] `http://toreopsahl.com/tnet/`

[5] We consider all subdomains in Yahoo (e.g., mail.yahoo.com), and other domains that belong to Yahoo (e.g., flickr.com).

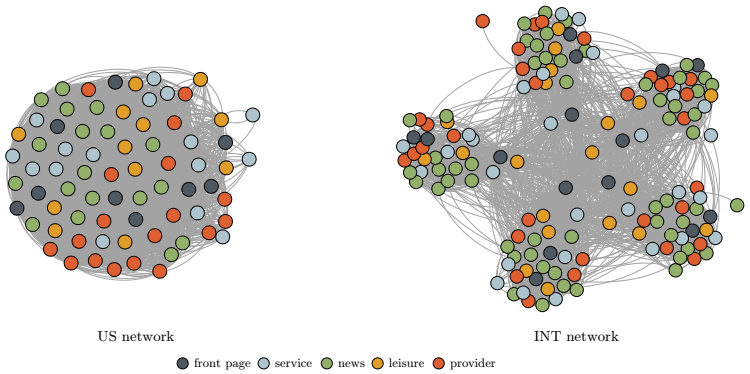


Figure 1: The US and INT provider networks based on browsing data of February 2014. The edges are not weighted for confidentiality reasons. The Force-Atlas algorithm is used for the visualisation.[6]

July 2014. The browsing data were used to create 365 networks ($INT_{01/08}$, ..., $INT_{31/07}$), each representing the traffic of the INT network on a certain day (here 08 refers to August, 07 to July, etc.). We also defined a network INT_{Feb} based on the browsing data of February 2014 to study the differences of inter-site engagement between sites. Finally, five country-based networks were created (INT_{c1},..,INT_{c5}), where each contains the sites and traffic of one country of the INT network. Looking at specific countries enables us to compare the inter-site engagement, on a per country basis.

In the remaining sections, unless otherwise started, we study the US network, more precisely, the network instances based on the browsing data of February 2014 (US_{Feb}), and the daily networks ($US_{01/08}$, ..., $US_{31/07}$).

Site categories. We annotated the sites in the networks using an adapted version of the schema described by Lehmann et al. (2013):

- 24 front pages and site maps (e.g., yahoo.com, everything.yahoo.com) [Front page]
- 46 service sites (e.g., mail.yahoo.com, calendar.yahoo.com) [Service]
- 73 news sites (e.g., news.yahoo.com, finance.yahoo.com) [News]
- 32 leisure and social media sites (e.g., tumblr.com, games.yahoo.com) [Leisure]
- 53 provider sites (e.g., account.yahoo.com, help.yahoo.com) [Provider]

Two examples of network instances are displayed in Figure 1. The nodes represent the sites colored by their category. On the left side, we can see the

[6] `https://gephi.github.io/`

very densely connected US network. The more central a node is the higher its connectivity to other nodes in the network. We notice that there is no site category that predominates the center in the network, i.e., each site category has densely and loosely connected nodes. On the right side, the INT network is displayed. We observe five densely connected modules, each representing a country. However, we can see connections between the modules implying that some users access sites from different countries.

3.2. Network-level metrics

Network-level metrics are concerned with the browsing behaviour within the whole provider network, which we refer to as the *provider engagement*. We employ three metrics of provider engagement. These are listed in Table 2, and can be of two types. The network *popularity* (first type) is measured by the number of sessions in which a site in the network was visited ($\#Sessions$). The browsing *activity* (second type) in the network during an online session is described by the average time users spend in the network ($DwellTime$) and the number of sites visited ($\#Sites$).

To measure the inter-site engagement, i.e., the traffic between sites, we employ standard graph metrics and add to them metrics that provide us with further information about the network structure. We refer to them as *inter-site* metrics. Numerous graph metrics exist. In this chapter, we focus on a subset of them, which are sufficient for our purpose. We discard metrics that could not be used to measure user engagement, and metrics for which we observed a very high correlation to those selected. This process resulted into the following five inter-site metrics. When describing them, we specify how they can be used in the context of user engagement.

Flow. The flow measures the extent to which users navigate between sites. It is defined as follows:

$$\frac{\sum_{i,j} w_{i,j}}{\sum_i v_i}$$

where $w_{i,j}$ is the total number of clicks between node n_i and n_j (i.e., the edge weight) and v_i is the total number of visits on node n_i. For example, a network with 6 visits, 3 on a service and 3 on a news site, can have different levels of flow. If there are 6 users, 3 solely visiting the service and 3 solely visiting the news site, the flow will be $0/6 = 0$. If two of the visits belong to one user accessing both sites, there will be traffic in the network, and the flow value will be $1/6$. A high value indicates a high inter-site engagement; users navigate often between sites in the network.

Density. The density (Wasserman, 1994) describes the connectivity of the network. It is the ratio between the number of edges compared to the number of all possible edges. In Figure 1 we see that the density of the US network is much higher than in the INT network, as there are few connections between nodes of different countries. A high connectivity (or density) means that the inter-site engagement is highly diverse; users navigate between many different sites.

Table 2: Network-level metrics: Metrics used to analyse the browsing behaviour within a provider network. $|N|$ refers to the number of nodes in the network.

Metric	Description	Engagement Low	Engagement High
Inter-site engagement			
Flow	Extent of the inter-site engagement.	0	$(\|N\|-1)/\|N\|$
Density	Diversity of inter-site engagement.	0	1
Reciprocity	Homogeneity of traffic between sites.	0	1
EntryDisparity	Variability of in-going traffic to the network.	1	0
ExitDisparity	Variability of out-going traffic from the network.	1	0
Provider engagement			
[POP] #Sessions	Total number of sessions in the network.	0	∞
[ACT] DwellTime	Avg. time per session in the network.	0	∞
[ACT] #Sites	Avg. number of sites per session in the network.	0	$\|N\|$

Reciprocity. The reciprocity measures the homogeneity of traffic between two sites, i.e., the percentage of traffic between two sites that is in both directions. We use the definition of Squartini et al. (2013) for the reciprocity of weighted networks:

$$\frac{\sum_{i<j} min[w_{i,j}, w_{j,i}]}{\sum_{i \neq j} w_{i,j}}$$

where $w_{i,j}$ is the weight of the edge from node n_i to node n_j. There are two reasons why a high reciprocity can be interpreted as a high engagement. First, the traffic from site i to j is from a different user group than the traffic from site j to i. In this case, a high reciprocity implies that the two user groups engage to the same extent on both sites. Second, the traffic between the sites comes from the same users. In this case, users do not only navigate from one site to another, they also return to the previously visited site.

Entry Disparity and Exit Disparity. The disparity refers to how the traffic to and from the network is distributed over the sites. For instance, a high entry (exit) disparity indicates that there are only few sites used to enter (leave) the network. We use the group degree measure of Freeman (1979) and adapt it:

$$\frac{\sum_i (g^*_{max} - g^*_i)}{|N| \cdot \sum_i g^*_i}$$

Here $|N|$ is the number of nodes in the network, g^{in}_i is the number of network visits that started at node n_i (user entered the network), and g^{out}_i is the number of network visits that ended after visiting node n_i (user left the network). The maximum values of g^{in}_i and g^{out}_i are defined by g^{in}_{max} and g^{out}_{max}, respectively.

We hypothesise that a low disparity (all nodes are equally used to enter and leave the network) reflects a high inter-site engagement. The network itself is less vulnerable, as the outage of one node (e.g., a front page) will not affect users entering the network. Moreover, it suggests that users do not need a front page to access other sites in the network; they know the site and go to it directly (maybe through a bookmark or a search site).

Table 3: Node-level metrics: Metrics used to analyse the browsing behaviour on a site in the provider network.

Metric	Definition	Engagement Low	Engagement High
Inter-site engagement			
PageRank	Probability that a user will visit the site.	0	1
Downstream	Avg. number of sites visited after visit on site.	0	∞
EntryProb	Probability that a user enters the network in this site.	0	1
ExitProb	Probability that a user leaves the network in this site.	1	0
Site engagement			
[POP] #Sessions	Total number of sessions on site.	0	∞
[ACT] DwellTime	Avg. time per session on site.	0	∞
Multitasking			
[MT] #Visits	Avg. number of visits per session on site.	0	∞
[MT] CumAct	Cumulative activity for the time between visits.	0	∞

3.3. Node-level metrics

Node-level metrics measure the browsing behaviour on a site within the network. Table 3 contains a list of all such metrics used in our work, their description, and their value range. We use two metrics that describe the engagement of users on the site (*site engagement*). The popularity of a site is measured by the number of sessions in which the site was visited ($\#Sessions$), and the browsing activity on a site is characterised by the average (median) time users spend on the site during a session ($DwellTime$). We additionally employ two *multitasking metrics* defined in Lehmann et al. (2013): the number of times a site was visited during an online session ($\#Visits$), and the cumulative activity ($CumAct$) which accounts for the absence time between visits within a session. Many visits and a long absence time between the visits is an indication of high loyalty to the site, i.e. the user is returning to the site to perform some new tasks within the same session (Lehmann et al., 2013).

We also employ four inter-site metrics, each accounting in a different way for the traffic to and from a site. All metrics consider the edge weights (number of clicks) between sites.

PageRank. The importance of pages in the Web is measured by PageRank (Page et al., 1999). The original definition considers the hyperlinks between pages, more precisely, the links leading to a page. Applied to our context, given the traffic between sites, $PageRank$ corresponds the probability that a user randomly navigating through the network will arrive at any particular site.

Downstream engagement. Whereas page rank measures the probability that a random user will visit any particular site, we analyse here the browsing behavior of a random user through the network who starts at a certain site. Motivated by Yom-Tov et al. (2013), we define downstream engagement in the context of traffic networks as follows. We use a discrete-time Markov process to simulate the browsing behaviour in the network. Hence, the sites correspond to the states and the edge weights correspond to the transition probabilities. Additionally, we assign to each site its exit probability (the definition is given later in this

section), i.e., at each step in the simulation there is a certain probability that a random user will leave the network.

For each site in the network, we now simulate the navigation of users through the network when starting on that site. The simulation ends when all users have left the network. Based on the simulated navigation paths of the users, we are able to compute several metrics; such as the time users spend in the network (i.e., sum of the dwell time of the visited sites), or the number of sites they visited (i.e., the path length). Each metric shows which sites are maximising the engagement to the network. Since the focus of our work is on interactions (traffic) between sites, we define downstream engagement as the average number of sites a random user visited according to the simulation.

Entry Probability and Exit Probability. The two metrics measure the probability that users enter or leave the network from the site under consideration. *EntryProb* is the percentage of visits to a site in which the user entered the network. *ExitProb* is the percentage of visits to a site from which the user leaves the network afterwards and thus does not continue browsing in the network. A high entry probability indicates that a site plays an important role in promoting inter-site engagement, whereas a high exit probability refers to a site with a negative effect on the inter-site engagement.

We next evaluate the value of these network- and node-level metrics in the context of user engagement, more precisely, in measuring inter-site engagement in a network of sites, such as those offered by Yahoo. Since our data do not follow a normal distribution, and thus to avoid the influence of heavy outliers, we do not use the metric values per site, but the corresponding ranks.

4. Evaluating inter-site metrics

We evaluate the applicability of inter-site metrics, defined in the previous section, to measure user engagement. The metrics at network-level enable us to compare provider networks, for instance, from different countries, showing, for instance, that some provider networks have a higher traffic than others. Additionally, the metrics at node-level enhance the understanding of how users engage with a single site and how the traffic between sites affects this. First, we compare inter-site metrics with standard engagement and multitasking metrics. Second, we present two case studies showing how inter-site metrics can be used to enhance our understanding of user engagement. We restrict ourselves to the INT network. This network is especially interesting, as it enables us to compare engagement between provider networks from different countries.[7]

4.1. Site and provider network rankings

We use the daily networks introduced in Section 3.1, namely $INT_{01/08}$, ..., $INT_{31/07}$. We calculate the network-level metrics for each network, and the

[7]Similar results were reached for the US network.

Table 4: Correlations between network-level metrics. Correlations with a p-value < 0.01 are not reported (-), and correlations above 0.5 or below -0.5 are highlighted in bold.

	Flow	Density	Reciprocity	EntryDisparity	ExitDisparity
Density	-				
Reciprocity	0.15	0.48			
EntryDisparity	0.23	**-0.61**	-0.38		
ExitDisparity	0.30	**-0.60**	-0.32	**0.84**	
[POP] #Sessions	-	**0.92**	0.42	**-0.54**	**-0.55**
[ACT] DwellTime	0.35	-0.45	-	0.33	0.38
[ACT] #Sites	**0.65**	-0.25	0.25	-	0.20

node-level metrics for all nodes in each network. We rank networks (nodes) according to each metric and then evaluate the similarity between these rankings using Spearman's rank correlation coefficient ρ. If two metrics produce the same ranking, they are equivalent and hence one is redundant. However, similar rankings may point to interesting dependencies between the engagement and traffic characteristics of a node or the whole network. We only report correlations that are statistically significant (p-value < 0.01).

Network-level metrics. The correlations between the network-level metrics are presented in Table 4. The density of a network is increasing (more sites become connected) with the number of sessions ($\rho = 0.92$). This means that the more users are visiting the network, the more diverse is the inter-site engagement; users visiting the network for many different reasons since they access different groups of sites. The metrics $Flow$ and $\#Sites$ are moderately correlated ($\rho = 0.65$). The more sites are visited during a session, the higher the flow of traffic. However, the correlation is not high enough to suggest that one of the metrics is redundant. Whether the traffic between two sites is unidirectional or not ($Reciprocity$) does not depend on any of the other considered metrics. We even cannot report a correlation to $DwellTime$, as the p-value is above 0.01.

Finally, we observe a strong correlation between the two disparity metrics, $EntryDisparity$ and $ExitDisparity$ ($\rho = 0.84$), indicating that the volume of in- and out-going traffic of the nodes depend on each other. The two metrics also correlate negatively with the density and the number of sessions of a network. This suggests that low engaging networks have some nodes that are used to enter (leave) the network (e.g., front pages), whereas in high engaging networks users enter (leave) the network over many nodes. However, both metrics are needed, as the correlations are only moderate.

To conclude, the metrics flow, reciprocity, and disparity capture distinct aspects of how users engage with a network of sites. The density, on the other hand, relates to the popularity of a network.

Node-level metrics. Table 5 reports the metric correlations at node-level. There are no correlations between the inter-site metrics, and the activity or multitasking metrics (all correlations are below 0.4). We therefore focus on the correlations between the inter-site metrics and the popularity metric $\#Sessions$.

We observe that popular sites in the provider network (e.g., front pages),

Table 5: Correlations between node-level metrics. Correlations above 0.5 or below -0.5 are highlighted in bold.

	PageRank	Downstream	EntryProb	ExitProb
Downstream	0.30			
EntryProb	-0.08	-0.27		
ExitProb	-0.10	-0.22	**0.79**	
[POP] #Sessions	**0.85**	0.17	0.12	0.08
[ACT] DwellTime	0.06	0.04	-0.19	-0.18
[MT] #Visits	0.08	0.02	0.13	0.18
[MT] CumAct	0.31	-0.02	0.35	0.32

are also visited frequently when browsing through the network ($\rho(\#Sessions, PageRank) = 0.85$), but users do not visit many other sites after visiting the site (we have $\rho(\#Sessions, Downstream) = 0.17$).

The strong correlation between $EntryProb$ and $ExitProb$ ($\rho = 0.79$) suggests that nodes used to enter the network are also frequently used to exit the network. The fact that these two metrics do not correlate with the other inter-site metrics indicates that entry and exit points of a network do not correspond to nodes that play an important role in directing traffic to other nodes (e.g., $\rho(EntryRatio, Downstream) = -0.27$), and to nodes that are visited frequently when browsing through the network (e.g., $\rho(EntryRatio, PageRank) = -0.10$).

In conclusion, downstream engagement, and the entry/exit ratio of a node bring new insights regarding the engagement at site level, whereas the page rank relates to sites that are very popular.

4.2. Case studies

We present two case studies that demonstrate how using inter-site metrics can enhance our understanding of user engagement. First, we compare different provider networks with respect to their inter-site engagement. Then, we use node-level metrics to analyse how the traffic differs between sites.

4.2.1. Comparing provider networks

The objective is to show that provider networks vary in their inter-site engagement. We compare the five country-based networks (INT_{c1}, ..., INT_{c5}) with each other. Using network-level metrics we compare engagement between networks, as we do it when comparing sites using node-level metrics. Figure 2 depicts the differences between the networks using four selected inter-site metrics. To capture the engagement to the provider network, we employ one provider engagement metric: $DwellTime$. The metrics are normalised by the z-score, hence the figure shows the extent to which the standard deviation of a metric rank is above or below the mean. The countries are ordered by decreasing $Flow$.

The highest inter-site and provider engagement can be observed for the first provider network (Country_1). The network has the highest flow, and dwell time. Also the reciprocity is above average. This shows that users spend a lot of time in that network, and while visiting the network they navigate often between many sites and do so in both directions. Although the network has the

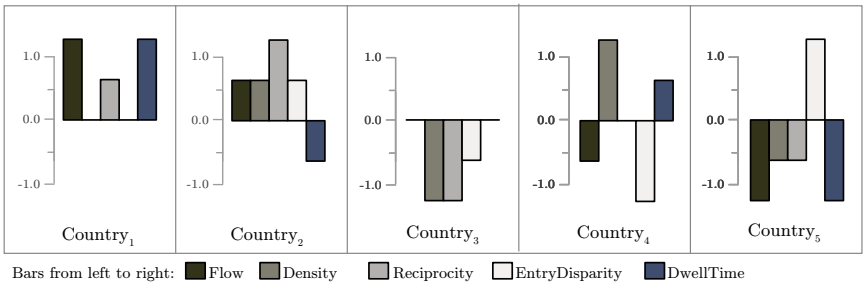


Figure 2: Comparing provider networks from different countries using network-level metrics.

highest engagement, the density is only average compared to the other networks, indicating that user do not navigate between many different sites.

We look now at the second provider network (Country_2). In this network, the flow is high and homogeneous (high *Flow* and *Reciprocity*), and also the diversity of the inter-site engagement is high (high *Density*), but the dwell time is below average (low *DwellTime*). This indicates that users access many sites in the network, but they navigate quickly between them.

The opposite can be observed for the fourth network (Country_4). The flow is below average, indicating a low inter-site engagement, but the dwell time per session is above average. We hypothesise that each user visits only a small subset of sites in the network, but spending a lot of time on it. The low value of *EntryDisparity* and the high value of *Density* suggests that still all sites in the network are visited, but from different users.

The last provider network (Country_5) has the lowest inter-site and provider engagement. We can see that users enter and leave the network over a subset of nodes (highest *EntryDisparity*). The users hardly navigate to other nodes (lowest *Flow* and low *Density*), or spend much time in the network (lowest *DwellTime*).

In this section, we demonstrated that network-level metrics enhance our understanding of how users engage with a whole provider network. Inter-site and provider engagement of a network can differ, implying that both metric types should be employed when analysing the engagement of a network of sites. Indeed, we saw that some networks have a high provider engagement, but a low inter-site engagement, and vice versa.

4.2.2. Engagement patterns at site-level

We study the different engagement patterns at site-level. Our goal is to show that inter-site metrics provide additional insights to those coming from assessing user behaviour within a site. The engagement patterns are determined based on several node-level metrics, *PageRank*, *Downstream*, *EntryProb*, *DwellTime*, and *CumAct* values. Each site is represented as a 5-nary vector, each dimension corresponding to one such metric. We cluster the vectors using k-means. The number of clusters is determined by a minimal cluster size such that each cluster

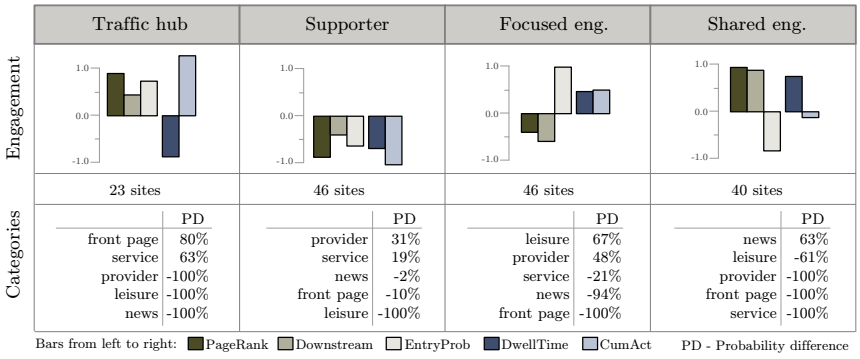


Figure 3: (1st row) Site clusters and browsing characteristics. (2nd row) Number of sites per cluster. (3rd row) Likelihood that a category occurs in a given cluster with respect to its likelihood that it occurs at all.

contains at least 20% of the sites under consideration (155 of them). We use the network constructed for February 2014 (INT_{Feb}).

We obtain four clusters, shown in Figure 3, each representing an engagement pattern. The first row contains the cluster centers normalised by the z-score, and the second row presents the number of sites within each cluster. The last row shows statistics related to the site categories per cluster. We define $p(c)$ as the probability that a site belongs to category c, and $p(c|cl)$ as the probability that a site in cluster cl belongs to category c. We then define the difference in the probability PD as follows:

$$\frac{p(c|cl) - p(c)}{max(p(c|cl), p(c))}$$

This corresponds to the likelihood that a category occurs in a given cluster with respect to its likelihood that it occurs at all. Each cluster, a pattern, is given a name reflecting its main characteristic.

Traffic hub. This cluster contains sites with a high inter-site engagement. The sites are important for the network, as they forward traffic to other sites. Users visit these sites when entering the provider network (high $EntryProb$) to access many other sites in the network (high $Downstream$). While browsing through the network, users regularly return to these sites, even after a long period of absence, to access further sites (high $PageRank$ and $CumAct$). We also observe the lowest $DwellTime$ for this cluster, which indicates that users do not spend much time on the sites; they quickly navigate to their target site. Front pages and service (e.g., search) sites belong to this cluster.

Supporter. Sites belonging to this cluster are sites on which users do not spend much time (low $DwellTime$), and do not return after a longer period of absence (low $CumAct$) during the session. The low $PageRank$ indicates that the sites are not very important (central) for the network, and hence they are not visited frequently. Provider (e.g., info.yahoo.com) and service (e.g., address.yahoo.com)

sites belong to this cluster. Users only visit the sites when specific information is required. As a result, users do not access these sites directly when entering the network (low $EntryProb$), but from other sites in the network. After they have found the required information, they are done, although they may visit a few other sites in the network (low $Downstream$).

Focused engagement. Many leisure sites belong to this cluster. We observe that users visiting leisure sites (game and social media sites) spend a lot of time on them (high $DwellTime$), and also return after a longer period of absence within the same session (high $CumAct$). It is also less likely that a user navigating through the network will arrive at a leisure site (low $PageRank$), and that a user that is visiting a leisure site will navigate to other sites in the network (low $Downstream$). Users access leisure sites directly (high $EntryProb$), and then they are solely engaged in their leisure activities. The same can be observed for some provider sites (e.g., messenger.yahoo.com), which are visited to download an application provided from Yahoo.

Shared engagement. Mainly news sites belong to this cluster, which is characterised by the highest inter-site engagement. Although the sites have the lowest entry probability, they are well connected to many other sites (highest $PageRank$). We hypothesise that users enter the network over front pages, and then visit a news site. Although they dwell long on the news site (high $DwellTime$), it is very likely that they continue browsing to other sites in the network (highest $Downstream$).

To summarise, we observe that sites exhibit different engagement patterns. For leisure sites, the attention of users is mostly focused on the site, whereas when reading news, users exhibit a high inter-site engagement (e.g., visiting several other sites). Then, there are sites that have the function to forward traffic to other sites (e.g., front pages) or to support users in the network (e.g., help sites).

5. Studying inter-site engagement

We study inter-site engagement with the metrics we have proposed and showed to bring different insights than standard engagement metrics. Our aim is to analyse how inter-site engagement differs depending on the loyalty of users, whether the user visits the network on a weekday or the weekend, and the upstream traffic. We also investigate the effect of users leaving the network, but returning within the same online session. We do so by defining further network instances and compare them with each other.

The network-level metrics $Flow$, $Density$, $Reciprocity$, $EntryDisparity$, $DwellTime$, and $\#Sites$ are used to characterise the inter-site engagement with respect to the whole provider network. The node-level metrics $PageRank$, $Downstream$, $EntryProb$, $DwellTime$, and $CumAct$ bring additional insights about the inter-site engagement with respect to a site. To study the difference between a metric value v_1 from one network with the metric value v_2 from

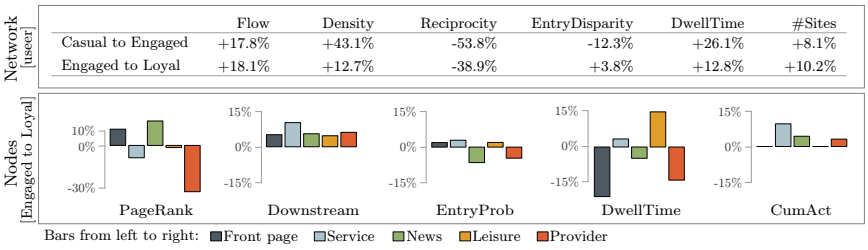


Figure 4: Differences between user-based networks.

another network, we measure the relative difference as:

$$d = \frac{v_2 - v_1}{max(v_1, v_2)}$$

where d is a value between -1 (decrease of -100%) and $+1$ (increase of $+100\%$).

The results for each type of subnetwork are presented in a Figure (e.g. Figure 4). The top part displays the differences of the network-level metrics, and the bottom part depicts the average differences of the node-level metrics per site category. The differences for each node-level metric are presented in a bar chart where a bar corresponds to a site category.

We start with comparing the network instances constructed to study how inter-site engagement is affected by user loyalty.

5.1. User Loyalty

We group users according to a loyalty criteria, measured in this chapter, by the number of active days within February 2014. Following Lehmann et al. (2012), we define three loyalty levels: Casual (1 active day), Engaged (2-14 active days), and Loyal (more than 14 active days). We then use the browsing data of February 2014 of the Casual, Engaged, and Loyal users to create three user-based networks.

We first analyse the differences at network level (Figure 4 (top part)). We see that Engaged users navigate more often (*Flow*: +17.8%), and between more sites (*Density*: +43.1%) than Casual users. The values increase again from Engaged to Loyal users. This shows that when the inter-site engagement increases, the more loyal the users are. Although we reported in Section 4.1 a weak positive correlation between the reciprocity and the density of a network ($\rho = 0.5$), we observe here that the reciprocity decreases with increasing loyalty of users (e.g., from Engaged to Loyal: -38.9%). This indicates that, with loyal users, the traffic in the network becomes more directed, more users go from one site to another but return less to the previous site. We speculate that, for instance, Engaged users always return to the front pages to access other sites in the network (e.g., $frontpage \rightarrow news \rightarrow frontpage \rightarrow leisure$). Loyal users, on the other hand, are "aware" of links that allow them to access other sites directly (e.g., $frontpage \rightarrow news \rightarrow leisure$). We also observe that the engagement on

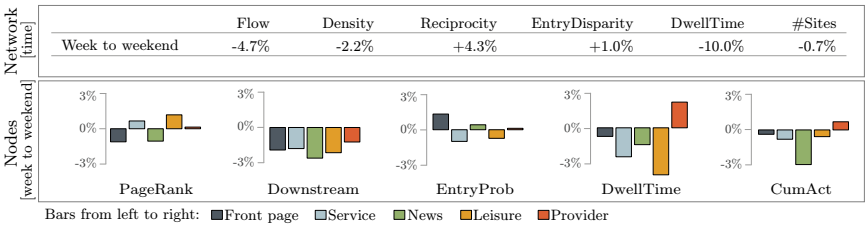


Figure 5: Differences between time-based networks.

the network increases with the loyalty of users. Loyal users spend more time on sites (*DwellTime* increases), and they also visit more sites (*#Sites* increases).

We analyse how the inter-site engagement differs at site level. Figure 4 (bottom part) compares the Engaged and Loyal networks. As already observed at network-level, loyal users browse more through the network. This is further accentuated at node-level by the increase in downstream engagement. For provider-related sites (e.g., help and account setting sites) we observe a significant decrease for page rank value, and also for the entry probability, and dwell time. This indicates that Loyal users are rarely visiting provider-related sites. They do not need to access help sites very much, likely because their account-related settings have already been performed a while ago.

Finally, in terms of time spent on the network, Loyal users seem to spend a considerable amount of time on leisure sites (increase in *DwellTime*), compared to Engaged users. We also observe a significant decrease of dwell time for front pages. Front pages can be compared to search sites; a low dwell time is a sign of a good user experience, as these sites are used to navigate (quickly) to other sites. We speculate that Loyal users know the front page well, and hence are able to move quickly to the site they want to reach. Engaged users, on the other hand, cannot find the hyperlink they are searching for immediately, and thus spend more time on such sites. As such, they may get more distracted by what is on offer on the front page.

5.2. Weekdays and weekend

Previous research showed that it is important to consider temporal aspects when studying user engagement (Lehmann et al., 2012). We therefore compare inter-site engagement during weekdays with that from weekends. We use the daily networks ($US_{01/08}$, ..., $US_{31/07}$) and split them depending on whether the network refers to traffic during a weekend or a weekday.

Although the differences are not as high as for the user-based networks, interesting observations can be made (see Figure 5). During the weekend, many metrics at network level (e.g., *Flow*: -4.7%, *DwellTime*: -10.0%), and site level (e.g., decrease in *Downstream* and *CumAct*) are lower. This means a lower site and inter-site engagement during the weekend. However, the reciprocity is higher (*Reciprocity*: +4.3%). We speculate that many users, who visit the network during the week, do it to perform specific goal-oriented tasks (e.g.,

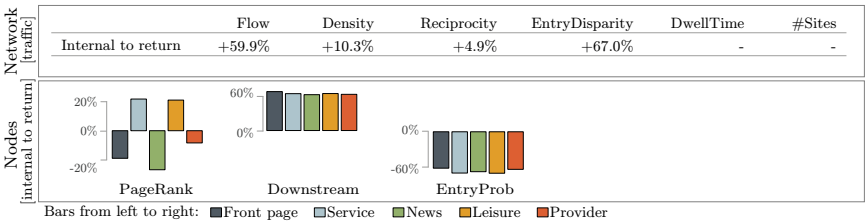


Figure 6: Differences between traffic-based networks.

checking mails, or reading news to keep up-to-date), and therefore navigate only from the front page to their desired site (i.e., the traffic is unidirectional).

The tasks during the weekend differ, as shown when looking at the statistics at node-level. During the weekend, users may not have to or do not wish to perform these goal-oriented tasks (lower $PageRank$ and $CumAct$ for front pages and news sites), and therefore they have the time to engage in leisure activities (higher $PageRank$), to do account-related settings and to try out applications offered by Yahoo (higher $DwellTime$ and $CumAct$ for provider sites).

5.3. Returning traffic

During their online sessions, users engage in multitasking (Lehmann et al., 2013), and as a result, they re-visit sites several times, after a short or long time with a same session. While doing so, users access sites outside the provider network, for instance, navigating from Yahoo mail to Facebook, and then back to Yahoo mail. In our data, we observed that on average 20% of the page views during an online session belong to sites that are not part of the provider network.

We analyse how this behaviour affects the characteristics of the networks. We therefore define a second type of edge, which we call "return edge", which corresponds to users navigating from a site in the provider network to external sites, but returning to another site in the network within the same online session (returning traffic). Figure 6 compares the internal traffic and the returning traffic of the US network of February 2014 (US_{Feb}), where for the latter network, return edges are added to the original network. Thereby, traffic returning to the same site n_i is represented by an additional edge $w_{i,i}$. Note that as our engagement and multitasking metrics do not consider the traffic between sites, their values are the same for the two networks. The reciprocity and closeness metrics do not consider traffic returning to the same site, as the two metrics are used to characterise the traffic between distinct sites. However, the metrics are still useful for analysing the change in the traffic in the network when accounting for returning traffic.

The results show that leaving the network does not necessarily entail less engagement. Users often return to the network ($Flow$: +59.9%) and more sites become connected through returning traffic ($Density$: +10.3%). A higher density indicates that users leave the network from one site and return to some other site in the network, but hardly ever navigate directly between the two

sites. This might point to missing hyperlinks in the provider network. Adding these hyperlinks could increase site and inter-site engagement, as they may help users browse through the network, and thus stay longer. How hyperlinks can influence the engagement in the network is investigated in Section 7.

The value of the entry disparity metric increases significantly (+67.0%), indicating that there are some sites that are less frequently used to enter the network when accounting for returning traffic; these sites are often used to return to the network within the same session. Interestingly, when looking at the entry probability per site category (node-level metrics), we are not able to identify a site category for which the entry probability decreases significantly more or less. The downstream engagement also increases to the same extent for all site categories. This suggests that whether the user is leaving the network and returning to it afterwards does not depend on the site category. In fact, the returning traffic is equally distributed over all categories of sites.

The only difference we can see is with respect to the *PageRank* metric. When we consider the returning traffic, the importance of service and leisure sites increases (higher *PageRank*), i.e., these sites become more connected through returning traffic

5.4. Upstream traffic

It was shown by Trevisiol et al. (2012) that user browsing behaviour within a network depends on where the user is coming from when entering the network (i.e., the upstream traffic). We investigate whether the upstream traffic has an effect on inter-site engagement. First, we define the upstream type (e.g., search, mail) of an online session. Using the referring URL and the schema from Lehmann et al. (2013), we annotate the sites from which users are coming from when entering the network. Additionally, we added the site category "Int" which refers to Yahoo sites that are not in the considered provider network (e.g., hk.yahoo.com when dealing with the US network). If the user accessed the network by using a bookmark, or entering the URL in the address bar, no referring URL is defined. In this case, the upstream type is "Tele", to refer to teleportation. This resulted into the following upstream types:

- 87.95% of teleportation [Tele]
- 4.32% of internal traffic (e.g., hk.yahoo.com, uk.news.yahoo.com) [Int]
- 1.42% of search (e.g., google.com, bing.com) [Search]
- 0.51% of social media (e.g., facebook.com, twitter.com) [Social]
- 0.19% of shopping (e.g., coupons.com, booking.com) [Shopping]
- 0.10% of news (e.g., cnn.com, forbes.com) [News]
- 0.07% of mail (e.g., live.com, mail.google.com) [Mail]

In total, 5.44% of the sessions could not be assigned with one of the defined upstream types, and are discarded in the following analysis. We now create a

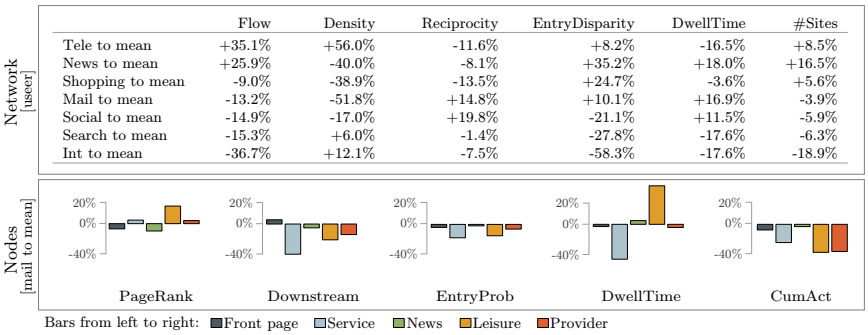


Figure 7: Differences between upstream-based networks.

network for each upstream traffic type, and calculate the network- and node-level metrics per network. Since there are many different types of upstream traffic, we do not compare the upstream traffic networks with each other. Instead, we compute the average value of each metric, and analyse the difference between the metric value and the average metric value.

Users frequently enter the network using teleportation (87.95% of the sessions). Teleportation is a sign that users are highly engaged with the network, as they use bookmarks, remember the domain name and enter it directly, or simply start typing the URL which then get autocompleted. This is also reflected by the network-level metrics in Figure 7 (top part) by the high values of density (+56.0%), traffic flow (+35.1%), and the average number of visited sites during a session (+8.5%). Interestingly, the dwell time in the provider network is below average (−16.5%).

A dwell time above average can be observed for users coming from news, mail, or social media sites. However, we can also see that the *Density* is below average, indicating that users do not navigate between many different sites. We know from Section 4.2.2 that news and social media sites inside the network also have a high dwell time. We speculate that users coming from such sites continue reading (socialising) on news (social media) sites inside the provider network. In doing so, they are highly engaged, as shown by the high value of dwell time. Users who come from external news sites (e.g., cnn.com) even visit many news sites inside the provider network (*Flow* and *#Sites* are above average).

We can see in the bottom part of Figure 7 (node-level metrics) that users who arrive from mail sites frequently visit leisure sites in the network, and that they spend a lot of time on them. The page rank and the dwell time is above average. Users might have received a notification via email from a leisure site, which led them to visit the site. We also observe that mail users focus much less on service sites, as all five metrics are below average (except *PageRank* which is on average).

The lowest engagement is observed for users who are coming from search or from other Yahoo sites. The flow and the two engagement metrics (*DwellTime* and *#Sites*) are low. However, the entry disparity is also below average (Search:

-27.8%, and Int: -58.3%), showing that users enter and leave the network from all sites. We speculate that users access directly the sites they are interested in (front pages are not used), to perform a quick task, and then leave again.

This section shows that inter-site engagement depends on many factors such as the loyalty of users and the day of the week. Accounting for multitasking (i.e., the returning traffic) also leads to a better understanding on how users engage with sites. In addition, considering where users are coming from provides information about what else the users are doing in the network afterwards. We have shown all these using metrics brought in to measure inter-site engagement.

6. The network effect

Liu et al. (2008) showed that the popularity of pages depends on the traffic between them, and we already observed in Section 4 that there is a strong correlation between site popularity and the inter-site engagement in the network. In this section, we demonstrate the effect of the network (the traffic between sites) on site engagement. We show that the traffic between sites affects the site popularity, and even slightly the activity on sites. Afterwards, we identify patterns that describe how the traffic is distributed over the network.

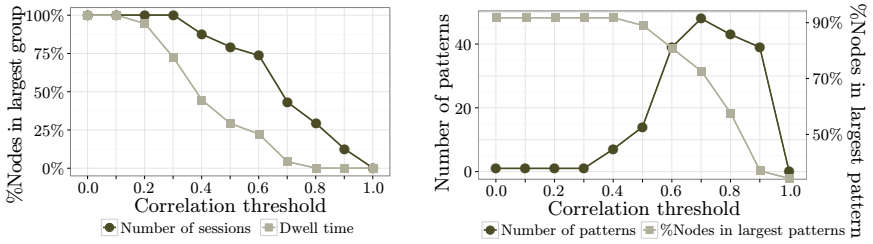
We use the daily US networks ($US_{01/08}$, ..., $US_{31/07}$) and removed networks modelling weekend browsing behaviours. This is to ensure that our observations are not caused by the difference in browsing behaviours between weekdays and weekends (see Section 5.2).

6.1. Dependencies between sites

We start by investigating the dependencies between sites in the provider network, to demonstrate the extent of the network effect. We want to see whether sites change their daily popularity (activity) in the same way. Based on the remaining 261 networks, we represent the daily popularity (activity) of a site by a vector $v_n = (c_1, ..., c_{261})$, where c_i is the number of sessions (average dwell time per session) on day i, for site n. We then compare the sites by calculating the Spearman rank correlation coefficient ρ between its vectors. Only statistically significant correlations are reported (p-value < 0.01). A high positive or negative correlation between two sites indicates that changes in the network are affecting both sites.

We study the strength of the network effect by grouping sites that are affected in the same way by changes in the network. We group all sites that have correlations above or below a given threshold θ. For instance, if $|\rho(a, b)| \geq \theta$, and $|\rho(a, c)| \geq \theta$, we create a group containing the sites $\{a, b, c\}$. We continued this process until all sites were compared with each other. Our results show that there is one large group, both in terms of popularity and activity of sites. Figure 8a shows the number of sites belonging to the largest group for increasing values of θ.

As expected, the size of the largest group decreases by increasing θ. However, when looking at the number of sessions, we still observe that 43.84% of the sites



(a) Network effect.

(b) Network effect patterns.

Figure 8: The strength of the network effect and corresponding patterns generated at different correlation thresholds.

affect each other with $\theta = 0.7$ (i.e., only considering correlations that satisfy $|\rho| \geq 0.7$). We can also report that only 2.93% of the correlations are negative ($\rho \leq -0.7$). This means that there is a significant positive network effect in terms of site popularity; sites become more *or* less popular together.

The effect on the activity metric $DwellTime$ is weaker. Only 14.67% of the sites belong to the largest group with $\theta = 0.7$. This implies that the activity on a site depends more on the site itself (e.g., users always spend more time on mail, but less on search). This was already observed in Lehmann et al. (2012). However, in this case 50% of the correlations are negative ($\rho \leq -0.7$). This shows that there are negative dependencies with respect to the time users spend on sites: an increase of dwell time on one site often leads to a decrease of dwell time on another site. We hypothesise that users have a limited amount of time to spent *on* the network *instead* of per site. Therefore, users switch from one site to another site (e.g., from Yahoo Sport to Yahoo Finance) within that limited time, thus more time spent on one site means less time spent on another site of the network.

6.2. Network effect patterns

We now study how the traffic between sites affects the site engagement. In other words, we analyse the spread of traffic through the network. We do so by looking at the dependencies between the edges as opposed to between the sites (as presented in the above section) to identify *network effect patterns* that describe how sites in the network exchange traffic with each other.

Similar to the first part in this section, we characterise the daily popularity of an edge by a vector $v_e = (c_1, ..., c_{261})$, where c_i is the number of clicks on day i, for edge e. We then compare the edges by calculating the Spearman rank correlations ρ between its vectors.[8] Finally, if the correlation is above or below a

[8] We excluded all edges with less than 30 clicks, to avoid the effect of minor fluctuations (e.g., [1,1,2] and [10,10,200] would have a correlation of 1). Using a threshold of 20 or 40 yields to similar results.

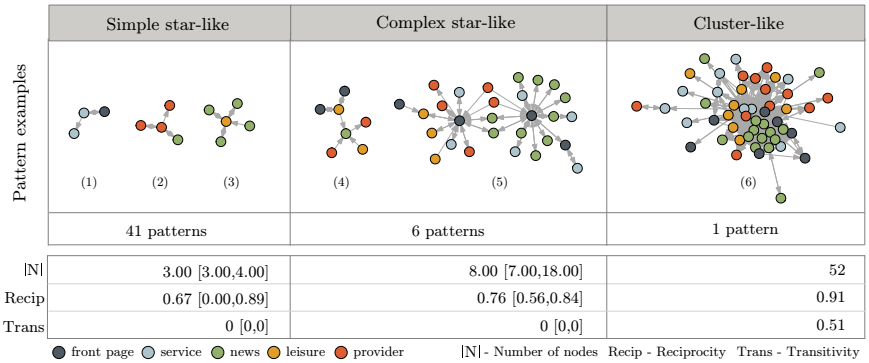


Figure 9: (1st row) Examples of network effect patterns. (2nd row) Number of patterns belonging to that group. (3rd row) Median of interquartile range of the number of sites ($|N|$), reciprocity ($Recip$), and transitivity ($Trans$) in each group.

given threshold θ, we says that the two edges are *related* in terms of the relative amount traffic passing by them.

Based on the resulting correlations, several patterns can be identified. If we observe a correlation $|\rho| \geq \theta$ between the edges $a \to b$ and $a \to c$, we create a "network effect pattern" containing the sites $\{a, b, c\}$ and the two edges. The pattern reflects that site a forwards traffic to site b and c. If there is also a correlation between $a \to c$ and $a \to d$, we add the site d and the edge $a \to d$ to that same "network effect pattern". We continue this process until all edges are compared with each other, and select all patterns that consist of at least three sites. We note that this approach enables us to analyse different network effect patterns involving the same site. For instance, if there is a correlation between $a \to b$ and $a \to c$, and $a \to d$ and $a \to e$, two "network effect patterns" can be observed. One pattern shows how site a forwards traffic to sites b and c, and the other one represents how the same site a directs traffic to sites d and e.

We study the number of patterns and the percentage of sites in the largest pattern for various threshold values θ (see Figure 8b). By increasing the threshold, the number of identified patterns increases, but the size of the largest pattern decreases. This means that parts of the largest pattern are divided into smaller patterns. We observe a peak of 48 patterns at $\theta = 0.7$. However, even with a threshold of $\theta = 0.7$, 71% of the sites are part of the largest pattern. We can conclude again that the network effect is significant. Changes in the network (e.g., increase of popularity of a site) affect many sites and the traffic between them, as shown by the largest network effect pattern. However, there are also smaller patterns that describe the network effect to smaller groups of sites. We can also report that the network effect is mainly positive (i.e., varies in the same direction), as only 1.3% of the correlations are below -0.7.

6.3. Examples of patterns

We focus on the patterns with correlations $|\rho| \geq 0.7$. We divide the patterns in three groups, shown in Figure 9. For each, we report the average number of sites, the average reciprocity, and the average transitivity (see Newman, 2003). We recall that the reciprocity describes the probability that the traffic between two sites flow in both directions, whereas the transitivity corresponds to the probability that two randomly selected neighbors of a site exchange traffic with each other. A low transitivity indicates that the pattern has a star-like structure; one site exchanges traffic with many other sites, and the other sites do not exchange traffic directly. A high transitivity reflects that all sites exchange traffic with each other. We refer to this as cluster-like structure.

The first two groups consist of network effect patterns with a star-like structures ($Trans = 0$). Simple star-like patterns have one focal site responsible for most traffic exchange, whereas complex star-like patterns have more than one focal site. In Figure 9 we see three examples of simple star-like (1,2,3) and two examples of complex star-like patterns (4,5). Surprisingly, other sites than front pages are responsible for the traffic exchange. In our examples, service (1), provider (2), leisure (3,4), and news sites (4) are focal sites.

In addition, focal sites are not necessarily connected with each other. In example (5), there are two provider sites that inject traffic to the focal sites (i.e., two edges of the type $\{provider\} \rightarrow \{frontpage\}$), and two news sites that exchange traffic with the focal sites (i.e., two edges of the type $\{news\} \leftrightarrow \{frontpage\}$). We also observe that the traffic often flows in both directions between two sites (e.g., complex star-like patterns have a median reciprocity of 0.76). This suggests that if a focal site increases the traffic to other sites, it is very likely that the other sites are also returning more traffic back.

On the right side of the figure, we see the largest network effect pattern, containing 52 sites in total. This pattern has a cluster-like structure; many sites exchange traffic directly with each other ($Trans = 0.52$), and also in both directions ($Recip = 0.91$). All site categories (e.g., mail, service, leisure) exchange traffic with each other.

In conclusion, the extent of the network effect in a provider network is significant; the traffic in the network affects the engagement of many sites. Next, we look at whether and how hyperlinks between the sites of a provider network influence this.

7. Hyperlink performance

Yom-Tov et al. (2013) demonstrated that hyperlinks help directing users to other sites in a provider network. Motivated by this, we analysed the different types of links on the sites of a provider network, and whether these influence the inter-site and site engagement.

For each site from the US provider network, we selected a random sample of pages accessed during February 2014 (US_{Feb}). We only considered pages that were accessed at least 10 times. In total, we sampled 43K pages. We

downloaded the HTML content of the pages, and extracted their hyperlinks.[9] We distinguished whether a link[10] points to a page within the provider network (*internal link*), or to somewhere else on the Web (*external link*). For the first case, we also differentiated between links to pages within the same site (*on-site links*), and links to pages to other sites (*inter-site links*) of the provider network. For each site, we then calculated the average percentage of on-site, inter-site, and external links per page. We did not consider the position and style of hyperlinks (we leave this for future work); our focus was the relationship between links and inter-site engagement.

7.1. Variations in the link structure

We first study whether sites differ in their hyperlink structure. Figure 10 shows the distribution of on-site, inter-site, and external links per site category. We report the median values.

Front pages have the highest percentage of inter-site links (62.1%). This is to be expected as they are used to access other sites in the network. However, the percentage of external links is also the highest compared to the other site categories (27.5%). A manual inspection shows that front pages are also used to direct users to sites outside the provider network. There are pages linking to Yahoo sites of other countries (e.g., `everything.yahoo.com/world`), or to sites of partnership providers (e.g., `att.yahoo.com`).

With service and news sites, on-site and inter-site links exist in the same frequency. Sites of both categories have around 40% on-site, and around 40% inter-site links. This results in 20% of external links.

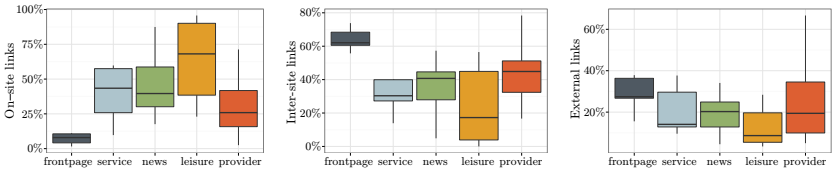
The highest on-site connectivity is given by leisure sites (68.11%). The on-site and inter-site hyperlink structure differs significantly among leisure sites. The interquartile range is between 38.5% and 90.1%, and 3.9% and 44.9%, respectively. Some leisure sites have many links between their pages, whereas others have many links to other sites in the provider network. However, all leisure sites do not link much to sites outside the provider network (8.7%).

Finally, provider sites do not have many in-site links, as many of them consists only of a few pages (e.g., info.yahoo.com). We observe as well that some provider sites have a high percentage of external links (the interquartile range is between 9.9% and 34.6%). These provider sites are also used from non-US users as an entry point to Yahoo (e.g., messenger.yahoo.com) and as such they link to the sites users are searching for (e.g., fr.messenger.yahoo.com for users from France).

Overall, this section shows variations in the link structure of a provider network, such as Yahoo US. We investigate next whether the link structure of a provider network has an effect on inter-site engagement.

[9] Pages from several sites were not considered, as signing in on the sites was required before downloading the pages.

[10] We use link and hyperlink interchangeably.



(a) On-site links. (b) Inter-site links. (c) External links.

Figure 10: Percentage of link types depending on site category.

Table 6: Spearman's correlation between site metrics using the link and traffic network. In case the p-value is above 0.01, we do not report the correlation (-).

(a) PageRank and downstream.

	Traffic	
	PageRank	Downstream
Hyperlinks		
PageRank	0.54	-
Downstream	-	-

(b) On-site, inter-site, and external.

	Traffic		
	On-site	Inter-site	External
Hyperlinks			
On-site	0.54	-0.45	-0.38
Inter-site	-0.40	0.50	-
External	-	-	0.39

7.2. Effect of the link structure

We investigate how the hyperlink structure of the provider network affects the browsing behavior of the users within the network. We model sites (nodes) and hyperlinks (edges) between them to form a *hyperlink network*. The edge weight is defined by the number of hyperlinks from one site to another.

We compare the hyperlink with the traffic network using the node-level metrics *PageRank* and *Downstream*.[11] Since the site engagement metrics cannot be employed on the hyperlink network, we also investigate how the composition of external, on-site and inter-site links of a site affects the browsing behavior of users when visiting that site. The browsing behaviour on a site in the traffic network is described by the average percentage of traffic to pages of the same site (*on-site traffic*), to other sites of the provider network (*inter-site traffic*), or to somewhere else (*external traffic*).[12] We then rank sites according to each measure in the traffic and hyperlink network and compare these rankings using the Spearman rank correlation coefficient ρ (p-value < 0.01). The results are presented in Table 6.

We observe in Table 6a that the importance of sites measured by *PageRank* is similar in both networks ($\rho = 0.54$). This suggests that if many hyperlinks lead to a certain site, it is also likely that users will visit that site. Interestingly, if a site has a high downstream engagement in the hyperlink network, it does not imply that users also navigate deeply into the network when visiting that site (p-value > 0.01).

[11] The exit probability in the hyperlink network is defined as the percentage of external links.
[12] The percentage of external traffic is the same as the node-level metric *ExitProb*.

In Table 6b, we can see that the likelihood that a user continues browsing within the same site depends on the percentage of on-site links ($\rho = 0.54$). At the same time, a high percentage of on-site links leads to less navigation between sites in the provider network, and to external sites ($\rho = -0.45$ and $\rho = -0.38$, respectively). On the other hand, sites with many links to other sites in the network have a high inter-site engagement ($\rho = 0.50$) and a low site engagement ($\rho = -0.40$); users navigate frequently to other sites in the network, but dwell less on the site under consideration.

External links do not influence the site and inter-site engagement, but we observe a weak correlation to the external traffic ($\rho = 0.39$). This suggests that providing external links leads to more users leaving the network. However, as we observed in Section 5.3, leaving the network does not necessarily imply less engagement, as users often return to the network within the same session.

In conclusion, whereas downstream engagement differs between the hyperlink and traffic network, the page rank applied to the hyperlink network can be used to identify sites that are also visited frequently by users. Moreover, providing more inter-site links and thus encouraging users to visit other sites in the provider network has a positive effect on inter-site engagement. These findings align with those reported in Yom-Tov et al. (2013). However, in doing so, the site engagement decreases which suggests that users only have a certain amount of time when being online. They use this time either mostly on one site, or across several sites within the network. This shows that there are dependencies between the inter-site and site engagement, and increasing both at the same time is a complex challenge.

8. Conclusions

This chapter proposed a methodology to study user engagement in a network of sites. We referred to this type of engagement as *inter-site engagement*. Large internet companies (e.g. AOL, Google, Yahoo) operate a network of sites, offering a variety of services, ranging from shopping to news. We model sites (nodes) and user traffic (edges) between them as a network, and employ metrics (at network- and node-level) from the area of complex graph analysis in conjunction with standard engagement metrics to study inter-site engagement. This enables us to analyse user engagement at a large-scale while accounting for the relationship between sites.

Five network-level and four node-level metrics were used to capture the inter-site engagement within the whole provider network and for individual sites, respectively. In addition to these graph metrics, we defined new metrics to study specific properties of inter-site engagement. For instance, we defined a flow metric to measure the extent to which users navigate between sites in the network. We also employ a metric that measures the downstream engagement, that is, how deeply users browse into the network after visiting a site. This metric enables us to identify sites that maximise the inter-site engagement. This is particularly important to know for front pages, as linking to them might

increase the engagement to the network. We referred to all these metrics as *inter-site* metrics. We demonstrated the value of these metrics by performing several case studies.

First, we compared inter-site metrics with standard engagement metrics. This brought various insights about inter-site engagement, which would not have been possible with standard engagement metrics alone. For instance, we observed that frequently visited sites lead users to access other sites in the provider network. Using network-level metrics, we showed that whole provider networks differ in their inter-site engagement. We could also identify networks that are highly engaging (users spend a lot of time in the network), but where the inter-site engagement is low (users do not visit many sites). In addition, we used node-level metrics to identify typical engagement patterns. We could show that users who visit the network for leisure activities stayed mostly on one site, whereas users interested in reading news visited several news sites.

We also analysed how returning traffic (users leaving the network but returning within the same session), user loyalty, and other aspects affect inter-site engagement. We saw that leaving the provider network does not necessarily entail less engagement, as many users return later on. As already observed in Lehmann et al. (2013), users often switch between sites and thereby access sites several times within an online session, effectively engaging in online multitasking. This suggests that providers should rethink about their "user engagement" strategy, which often comes down to keeping users as long as possible on their sites. Instead, it may be beneficial (long-term) to entice users to leave the network (e.g. by offering them interesting off-network content in the context of news sites) in a way that users will want to return to it (e.g., become a reference site).

We investigated the dependencies between site engagement and traffic between sites in the provider network, which we call the *network effect*. We showed that there is a strong network effect with respect to site popularity, i.e., changes in the network affect the traffic (on edges) and hence many sites in the network. Although the activity on a site depends more on the site itself, we still observed that an increase in activity of one site can lead to a decrease in activity on another site. This suggests that users will very often only have a limited amount of time when online. If they have to visit several sites on the network, they are likely to do so quickly, so that to keep within their available time.

Finally, we compared the traffic and hyperlink network with each other, and showed that hyperlinks can influence the user browsing behavior in the network. Whereas the downstream engagement in the two networks did not align, the importance of sites measured by *PageRank* is similar in both networks; if many hyperlinks lead to a certain site, it is also likely that users will visit that site. We also found that more hyperlinks between sites lead to a higher inter-site engagement (users access more sites), but to a lower engagement on sites (users spend less time on sites). This means that site and inter-site engagement influence each other, and improving both at the same time may be difficult.

Future work. Several lines of research can be pursued. We looked at the

effect of several dimensions on inter-site engagement separately. An important extension of our work will be to combine these dimensions, for instance, studying the behaviour of loyal users in each country. Other dimensions should also be considered, including demographics and in particular finer grained time analysis, such as morning versus evening.

We also did not differentiate between the ways users navigate between sites, whether clicking on hyperlinks, using browser tabs, or bookmarks. We could build additional types of traffic networks that, for instance, account only for the traffic produced by clicking on hyperlinks and compare it with the hyperlink network. This can bring new insights into how hyperlinks influence inter-site engagement, and which are the hyperlinks that are important in directing traffic to other sites. In this context, one could also analyse how the position and style of hyperlinks contribute to this.

Acknowledgements. This work was partially funded by Grant TIN2012-38741 (Understanding Social Media: An Integrated Data Mining Approach) of the Ministry of Economy and Competitiveness of Spain. This work was carried out as part of Janette Lehmann's PhD internship at Yahoo Labs Barcelona. We thank the Yahoo Toolbar users for providing their browsing log data.

References

Barabási, A.-L., Albert, R., Jeong, H., 2000. Scale-free characteristics of random networks: The topology of the world-wide web. Physica A: Statistical Mechanics and its Applications. 281 (1), 69–77.

Beauvisage, T., 2009. The dynamics of personal territories on the web. In: Proceedings of the 20th ACM Conference on Hypertext and Hypermedia (HT). pp. 25–34.

Bucklin, R. E., Sismeiro, C., 2003. A model of web site browsing behavior estimated on clickstream data. Marketing Research. 40 (3), 249–267.

Catledge, L. D., Pitkow, J. E., 1995. Characterizing browsing strategies in the world-wide web. Computer Networks and ISDN systems. 27 (6), 1065–1073.

Chen, W., Wang, Y., Yang, S., 2009. Efficient influence maximization in social networks. In: Proceedings of the 15th ACM Conference on Knowledge Discovery and Data Mining (SIGKDD). pp. 199–208.

Cheng, S., Shen, H., Huang, J., Zhang, G., Cheng, X., 2013. Staticgreedy: Solving the scalability-accuracy dilemma in influence maximization. In: Proceedings of the 22nd ACM Conference on Information and Knowledge Management (CIKM). pp. 509–518.

Chmiel, A., Kowalska, K., Hołyst, J. A., 2009. Scaling of human behavior during portal browsing. Physical Review E. 80 (6), 066122.

Cockburn, A., McKenzie, B., 2001. What do web users do? an empirical analysis of web use. International Journal of Human-computer Studies. 54 (6), 903–922.

De Maeyer, J., 2011. Hyperlinks and journalism: Where do they connect? In: Future of Journalism Conference.

Dellarocas, C., Katona, Z., Rand, W., 2013. Media, aggregators, and the link economy: Strategic hyperlink formation in content networks. Management Science. 59 (10), 2360–2379.

Dupret, G., Lalmas, M., 2013. Absence time and user engagement: Evaluating ranking functions. In: Proceedings of the 6th ACM Conference on Web Search and Data Mining (WSDM). pp. 173–182.

Freeman, L. C., 1979. Centrality in social networks conceptual clarification. Social Networks. 1 (3), 215–239.

Jiang, Q., Tan, C.-H., Wei, K.-K., 2012. Cross-website navigation behavior and purchase commitment: A pluralistic field research. In: Proceedings of the 16th Pacific Asia Conference on Information Systems (PACIS). p. 193.

Johnson, E. J., Moe, W. W., Fader, P. S., Bellman, S., Lohse, G. L., 2004. On the depth and dynamics of online search behavior. Management Science. 50 (3), 299–308.

Kempe, D., Kleinberg, J., Tardos, É., 2003. Maximizing the spread of influence through a social network. In: Proceedings of the 9th ACM Conference on Knowledge Discovery and Data Mining (SIGKDD). pp. 137–146.

Koidl, K., Conlan, O., Wade, V., 2014. Cross-site personalization: Assisting users in addressing information needs that span independently hosted websites. In: Proceedings of the 25th ACM Conference on Hypertext and Hypermedia (HT). pp. 66–76.

Kumar, R., Tomkins, A., 2010. A characterization of online browsing behavior. In: Proceedings of the 19th Conference on World Wide Web (WWW). pp. 561–570.

Lehmann, J., Lalmas, M., Dupret, G., Baeza-Yates, R. A., 2013. Online multitasking and user engagement. In: Proceedings of the 22nd ACM Conference on Information and Knowledge Management (CIKM). pp. 519–528.

Lehmann, J., Lalmas, M., Yom-Tov, E., Dupret, G., 2012. Models of user engagement. In: Proceedings of the 20th Conference on User Modeling, Adaptation, and Personalization (UMAP). Springer, pp. 164–175.

Liu, Y., Gao, B., Liu, T.-Y., Zhang, Y., Ma, Z., He, S., Li, H., 2008. Browserank: Letting web users vote for page importance. In: Proceedings of the 31st ACM Conference on Research and Development in Information Retrieval (SIGIR). pp. 451–458.

Meiss, M. R., Gonçalves, B., Ramasco, J. J., Flammini, A., Menczer, F., 2010. Agents, bookmarks and clicks: A topical model of web navigation. In: Proceedings of the 21st ACM Conference on Hypertext and Hypermedia (HT). pp. 229–234.

Meiss, M. R., Menczer, F., Fortunato, S., Flammini, A., Vespignani, A., 2008. Ranking web sites with real user traffic. In: Proceedings of the 1st ACM Conference on Web Search and Data Mining (WSDM). pp. 65–76.

Newman, M. E., 2003. The structure and function of complex networks. SIAM review. 45 (2), 167–256.

O'Brien, H. L., Toms, E. G., 2008. What is user engagement? A conceptual framework for defining user engagement with technology. Journal of the American Society for Information Science and Technology (JASIS). 59 (6), 938–955.

Page, L., Brin, S., Motwani, R., Winograd, T., 1999. The pagerank citation ranking: Bringing order to the web. Tech. rep., Stanford InfoLab.

Park, Y.-H., Fader, P. S., 2004. Modeling browsing behavior at multiple websites. Marketing Science. 23 (3), 280–303.

Peterson, E. T., Carrabis, J., 2008. Measuring the immeasurable: Visitor engagement. Tech. rep., Web Analytics Demystified.

Roos, J. M., 2012. Hyper-media search and consumption. Ph.D. thesis, Duke University.

Simkin, M., Roychowdhury, V., 2008. A theory of web traffic. Europhysics Letters (EPL). 82 (2), 28006.

Squartini, T., Picciolo, F., Ruzzenenti, F., Garlaschelli, D., 2013. Reciprocity of weighted networks. Scientific reports. 3.

The PEW Research Center, 2010. Understanding the participatory news consumer. `http://www.pewinternet.org/~/media/Files/Reports/2010/PIP_Understanding_the_Participatory_News_Consumer.pdf`.

The PEW Research Center, 2012. In changing news landscape, even television is vulnerable. `http://www.people-press.org/files/legacy-pdf/2012NewsConsumptionReport.pdf`.

Trevisiol, M., Aiello, L. M., Schifanella, R., Jaimes, A., 2014. Cold-start news recommendation with domain-dependent browse graph. In: Proceedings of the 8th ACM Conference on Recommender Systems (RecSys). pp. 81–88.

Trevisiol, M., Chiarandini, L., Aiello, L. M., Jaimes, A., 2012. Image ranking based on user browsing behavior. In: Proceedings of the 35th ACM Conference on Research and Development in Information Retrieval (SIGIR). pp. 445–454.

Wasserman, S., 1994. Social network analysis: Methods and applications. Vol. 8. Cambridge University Press.

Wu, X., Yu, K., Wang, X., 2011. On the growth of internet application flows: A complex network perspective. In: Proceedings of the 30th IEEE Conference on Computer Communications (INFOCOM). pp. 2096–2104.

Xue, W., Shi, J., Yang, B., 2010. X-rime: Cloud-based large scale social network analysis. In: Proceedings of the 7th IEEE Conference on Services Computing (SCC). pp. 506–513.

Yom-Tov, E., Lalmas, M., Baeza-Yates, R., Dupret, G., Lehmann, J., Donmez, P., 2013. Measuring inter-site engagement. In: Proceedings of the IEEE Conference on Big Data (BigData). pp. 228–236.