

From Site to Inter-site User Engagement: Fundamentals and Applications

Janette Lehmann

PhD Thesis / 2014

Supervisors:

Prof. Dr. Ricardo Baeza-Yates

Department of Information and Communication Technologies

Dr. Mounia Lalmas

Yahoo Labs





Acknowledgements

Don't panic, was probably one of my first thoughts when I started my PhD, and it was not the last time that I had this thought. However, although I had to face demanding challenges, it was definitely worth it. In this context, I want to thank many people that guided and supported me during this exciting time.

First of all, I want to express my sincerest gratitude to my supervisors, Prof. Dr. Ricardo Baeza-Yates and Dr. Mounia Lalmas, who accompanied me the last years with their knowledge and advice, and made it possible for me to work on a research field that captivated me totally. I am very grateful to Ricardo for giving me the opportunity to do a PhD in such a great research environment, for always having time and being helpful, and for sharing his deep knowledge with me. I am deeply thankful to Mounia for supporting me at any time, for bringing me back to a more global view when I was lost in the data, and especially for teaching me how to write research papers – which, of course, was not always easy for her. Without their guidance this thesis would have not been possible.

Looking back, I also want to thank Prof. Dr. Claudia Müller-Birn for introducing me to the world of research. If she would not have approached me, seeing my potentials, and fascinating me with her passion about research and life, I would never have gone in this direction. Without her, I also would never have met Dr. Ciro Cattuto, and worked at the ISI Foundation in Turin – my first step into a life as a full-time researcher.

Ciro was the second person I met that had deep influence on my further research career. I want to thank him for many inspiring discussions – al-

though his time is always limited, since he is travelling so much that he probably makes half the way to the moon every year. I am grateful to have worked with him, especially when he shared his excitement about technology, research, and many other things with me. *Ciro* was also the person that encouraged me to go to Barcelona to meet *Ricardo* and to introduce myself for a PhD position.

There were many others that had a formative influence on me during the last years. From the beginning of my PhD, I especially remember *Georges Dupret* and *Elad Yom-Tov*, who helped me solving tricky data mining and algorithmic problems. I will never forget the Wikipedia project with *David Laniado* and *Andreas Kaltenbrunner*. Thank you so much for being part of the team from the beginning until the very end.

I will also never forget the lessons I have learned during my two internships. I am deeply thankful to *Carlos Castillo* for inviting me to Qatar, and teaching me how to become a well-organised and efficient researcher. Just before finishing the PhD, I went to London for another internship. Apart from the fact that all my colleagues there formed an amazing team, I especially want to thank *John Agapiou* for sitting down with me for hours and trying to understand the data with his statistical master mind when I could not see a light at the end of the tunnel anymore, and *Andriy Bychay* for helping me with all the cluster and account-related issues.

During the very last days of my PhD, I had support from another great group of people – *Bruce*, *Christian*, *Gianmarco*, *Hugues*, *Julie*, *Julio*, *Jordi*, *Lydia* – thank you for all the proof-readings, translations, and for managing some administrative and organisational issues that I could not handle as I was not in Barcelona.

I am deeply grateful to many people in and outside the lab for making my work and life in Barcelona as it was. First of all, *Michele Trevisol* for being such a great “desk mate” with his cheerful and positive attitude, and also for always having the time for a break (???) and keeping me up-to-date about the latest gossip. I am also grateful to many other colleagues who made my life at work so much enjoyable and memorable, especially *Cidgem*, *Gianmarco*, my dear *Luca*, and *Miriam*. I am also deeply thankful that I could share the ups and downs of a PhD with such great persons like *Ruth*, *Luca*, and *Eduardo*. Thank you, *Natalia* and *Estefania*, for making sure that everything was working smoothly in the lab, and also that we always had at least one working coffee machine.

I would like to thank my flatmates Ronaldo, Julio, Bruce, Charles, and Nevat for giving me a home and being my family in Barcelona, and my dear friends Julie and Hugues for nights of inspiring conversations, great wine and spicy food. And all the others that I have met in Barcelona, such as Ahmet, Audi, Manos, Soroosh, and the Ragazzi di Calle Ample for giving me the right balance between work and life.

Beside the people already mentioned, there is a long list of people that gave me strength and supported me remotely. First, my beloved partner Kay, who had the patience for a long-distance relationship with me, and who has not gone mad the last months, although I totally occupied his flat with documents, papers, diagrams, and myself. Who, although writing now as well, cheered me up when I lost my confidence, but also left me alone in my “world of thoughts” when necessary.

I would like to thank my family for all their love and trust. Especially my mum, who is always there for me and always believes in me, and my father from whom I have got my calm and balanced personality – something that, I believe, is very helpful for handling all the challenges of life. My brother, a person like my father, always endearing and helpful, and my great-grandmother who believes she is old (which is true), but not as old as she thinks. You all shaped me since my childhood.

My thanks also go to my friends from around the world: Karo the toughest women I have ever met, always telling me what I have to hear, but not what I want to hear, Cola for being the best friend in my life, and Sassi for being my oldest friend and a cuddling and dancing bear. And also to Ari and Fede my hearts of Turin, and many others: Agi, Burkhard, Hanni, Kagi, Rowena, Suse, Young-Ho, Zoltan, ... The list is very long, and I feel honoured to have you in my life. We all went together through so many good, bad and crazy times, that I will never forget.

Thank you all for understanding that my time was limited the last years, thank you for believing in me, and thank you for visiting me wherever I was.

All these people and many others, their support and understanding, and my passion for getting to the bottom of the things, have brought me so far.

So Long, and Thanks For All the Fish. - Douglas Adams.



Abstract

The aim of this thesis is to provide a deeper understanding of how users engage with websites and how to measure this engagement. We start with studying online behaviour metrics, which are commonly employed as a *proxy* for user engagement, and we propose new metrics that expose so far unconsidered aspects of user engagement. We then conduct several case studies that demonstrate how these metrics provide a deeper understanding of user engagement. Within each case study we also examine how the characteristics of a website influence user engagement.

Some of our key findings include: (1) engagement differs between sites and these differences depend on the site itself; (2) users multitask within on-line sessions and this affects the interpretation of engagement metrics; (3) analysing engagement across sites enables a comprehensive look at user engagement, because this considers the relationships between sites; and (4) engagement depends on the quality of the content and the hyperlink structure of sites, but the interests of users can also drive it.

Resum

L'objectiu d'aquesta tesi és aprofundir en el concepte de participació i compromís dels usuaris en pàgines web i analitzar com mesurar aquesta interacció. Comencem estudiant les mètriques del comportament online, les quals s'utilitzen habitualment com a representació de la participació de l'usuari, i proposem noves mètriques que consideren aspectes inexplorats de la participació i compromís de l'usuari. A continuació, analitzem una sèrie de casos d'estudi que demostren com aquestes mètriques proporcionen una millor comprensió de la participació i el compromís dels usuaris. En cadascun d'aquests casos d'estudi també analitzem la manera amb la qual les característiques de la pàgina web poden influenciar aquesta interacció.

Les nostres troballes principals són: (1) la participació i el compromís de l'usuari varia entre pàgines web i aquesta depèn de les característiques de les mateixes; (2) els usuaris realitzen diverses tasques simultàniament dins les sessions online i això influeix en la interpretació de les mètriques considerades; (3) analitzar la participació i el compromís de l'usuari en diferents pàgines web permet obtenir una comprensió global de les relacions entre elles; i (4) la participació i compromís de l'usuari depèn de la qualitat del contingut i de l'estructura d'enllaços de la pàgina, però també es fonamenta en els propis interessos de l'usuari.

Resumen

El objetivo de esta tesis es profundizar en el concepto de participación y compromiso de los usuarios en sitios web y analizar cómo medir dicha interacción. Empezamos estudiando las métricas del comportamiento *online*, las cuales se utilizan habitualmente como representación de la participación del usuario, y proponemos nuevas métricas que consideran aspectos inexplorados de la participación y compromiso del usuario. A continuación, analizamos varios casos de estudio que demuestran cómo estas métricas proporcionan una mejor comprensión de la participación y compromiso de los usuarios. En cada uno de estos casos de estudio también analizamos la manera con la que las características del sitio web pueden influenciar esta interacción.

Nuestros principales descubrimientos son: (1) la participación y el compromiso del usuario varía entre sitios web y ésta depende de las características de las mismas; (2) los usuarios realizan varias tareas simultáneamente en las sesiones online y esto influye en la interpretación de las métricas consideradas; (3) analizar la participación y el compromiso del usuario en diferentes sitios web permite obtener una comprensión global de las relaciones entre ellas; y (4) la participación y compromiso del usuario depende de la calidad del contenido y de la estructura de enlaces del sitio, pero también puede deberse los intereses de los usuarios.



Contents

Abstract	vii
Resum	viii
Resumen	ix
Contents	xi
List of Figures	xvi
List of Tables	xix
1 Introduction	1
1.1 Research Problems	2
1.2 Thesis Structure and Main Contributions	5
I User Engagement	11
2 Background	13
2.1 Definition of User Engagement	13
2.2 Measurements	14
2.3 Factors of Influence	17
3 Methodology	21
3.1 Interaction Data	21
3.2 Online Sessions and Site Visits	22

3.3	Website Taxonomy	24
II Fundamentals		25
4	Site Engagement	27
4.1	Introduction	27
4.2	Dataset	28
4.3	Site Engagement Metrics	29
4.4	Diversity in Engagement	30
4.5	Patterns of Site Engagement	36
4.5.1	General Patterns	36
4.5.2	User-based Patterns	38
4.5.3	Time-based Patterns	39
4.6	Relationship between Patterns	40
4.7	Discussion	42
5	Online Multitasking	45
5.1	Introduction	45
5.2	Related Work	47
5.3	Dataset	49
5.4	Navigation Model	49
5.5	Characteristics and Effects	53
5.5.1	Multitasking in Sessions	53
5.5.2	Visit and Inter-visit Activity	55
5.6	Multitasking Metrics	58
5.6.1	Cumulative Activity	59
5.6.2	Activity Patterns	61
5.7	Evaluation	64
5.8	Multitasking Patterns	66
5.9	Relationship between Patterns	69
5.10	Discussion	71
6	Inter-site Engagement	75
6.1	Introduction	75
6.2	Related Work	76
6.3	Dataset and Networks	78
6.4	Characteristics and Effects	80
6.4.1	Traffic and Popularity	80
6.4.2	Entry and Exit Points	82

6.5	Inter-site Engagement Metrics	83
6.5.1	Network-level Metrics	83
6.5.2	Node-level Metrics	86
6.6	Evaluation	88
6.7	Comparing Provider Networks	90
6.8	Patterns of Inter-site Engagement	92
6.9	Discussion	94
 III Applications I: Site Engagement		97
7	Native Advertising	99
7.1	Introduction	99
7.2	Related Work	102
7.3	Measuring Post-click Experience	104
7.4	Effect on User Engagement	106
7.5	Differences between Mobile and Desktop	109
7.6	Predicting High Quality Ads	112
7.6.1	Problem Definition and Methods	113
7.6.2	Offline Models Evaluation	115
7.6.3	Feature Ranking	116
7.7	Online Bucketing Evaluation	117
7.8	Discussion	122
8	Reader Engagement in Wikipedia	125
8.1	Introduction	125
8.2	Related Work	127
8.3	Datasets	129
8.4	Reading Preference	131
8.4.1	Popular Topics	131
8.4.2	Reading and Editing Preferences	133
8.4.3	Preference Matrix	134
8.5	Reading Behaviour	137
8.5.1	Reading Patterns	138
8.5.2	Changes in Reading Patterns	142
8.6	Discussion	145

IV Applications II: Inter-Site Engagement	151
9 Engagement in a Provider Network	153
9.1 Introduction	153
9.2 Dataset and Networks	154
9.3 Diversity in Inter-site Engagement	155
9.3.1 User Loyalty	156
9.3.2 Weekdays versus Weekend	157
9.3.3 Returning Traffic	158
9.3.4 Upstream Traffic	160
9.4 The Network Effect	162
9.4.1 Dependencies between Sites	162
9.4.2 Network Effect Patterns	164
9.5 Hyperlink Performance	167
9.6 Discussion	170
10 Story-Focused Reading across News Sites	173
10.1 Introduction	173
10.2 Related Work	175
10.3 Dataset	177
10.4 Story-Focused News Reading	180
10.4.1 Shuffle Test	180
10.4.2 Popularity and Providers	182
10.5 Users and News Sections	183
10.5.1 Story-Focused Reading and Users	183
10.5.2 Story-Focused Reading and Sections	184
10.6 Session Characteristics	185
10.6.1 Focused versus Non-focused Sessions	186
10.6.2 Depth of Story-focused Reading	188
10.6.3 Upstream Traffic	189
10.7 Hyperlink Performance	192
10.7.1 Number of Inline Links	193
10.7.2 Types of Inline Links	194
10.7.3 Position of Inline Links	196
10.8 Effect on User Engagement	197
10.9 Discussion	200
11 Discovering Story-related Content in Twitter	203
11.1 Introduction	203
11.2 Related Work	206

11.3	Datasets	208
11.4	Crowd Characteristics	211
11.4.1	Crowd Creation	211
11.4.2	Crowd Members	211
11.4.3	Crowd Dynamics	214
11.5	Crowd-based News Discovery	216
11.5.1	Candidate Generation	217
11.5.2	Learning Framework	219
11.5.3	Training Data	219
11.5.4	Features	220
11.5.5	Results	223
11.5.6	Application	224
11.6	News Story Curators	225
11.6.1	Concepts	225
11.6.2	Learning Framework	228
11.6.3	Training Data	229
11.6.4	Features	230
11.6.5	Results	231
11.6.6	Precision-oriented Evaluation	232
11.7	Discussion	233
12 Conclusions and Future Work		237
12.1	Summary	237
12.2	Research Results	240
12.3	Future Work	247
Bibliography		249
Appendix		275
A.1	Predicting High Quality Ads	275
A.1.1	Ad Landing Page Features	275
A.1.2	Offline Models Evaluation	277
A.2	Story-focused Reading	279
A.3	Crowd-based News and Curator Discovery	279
A.3.1	Labelling News Articles	279
A.3.2	Labelling News Story Curators	280

List of Figures

1.1	Structure of the thesis.	6
3.1	Example of interaction data.	23
4.1	Normalised engagement values per site.	30
4.2	Site clusters and user groups.	33
4.3	Engagement over time using $\#Users$	35
4.4	General patterns of engagement.	37
4.5	User-based patterns of engagement.	38
4.6	Time-based patterns of engagement.	40
5.1	From click-streams to tree-streams.	50
5.2	Site visit characteristics for four categories of sites.	56
5.3	Activity patterns during online sessions.	57
5.4	Cumulative activity for different importance values of the absence times.	61
5.5	Activity patterns for different numbers of visits in a session.	63
5.6	Multitasking patterns during online sessions.	67
5.7	Distribution of sites of a category over the clusters.	70
6.1	The provider network based on browsing data of February 2014.	79
6.2	Fluctuations of network popularity and traffic over time.	80
6.3	Distributions of node weights (popularity) and edge weights (traffic).	81
6.4	Network characteristics.	82
6.5	Comparing provider networks from different countries.	91

6.6	Inter-site engagement patterns.	93
7.1	Example of a native ad (second item) in a news stream on a mobile device.	100
7.2	Post-click experience. The probability of a second click given dwell time.	105
7.3	Changes in engagement depending on whether the users experienced short or long clicks.	107
7.4	Distributions of the differences in post-click experience between mobile and desktop.	109
7.5	Example of a landing page (mobile optimised vs. non-optimised).	110
7.6	Differences in post-click experience between mobile and desktop depending on whether a landing page is mobile optimised or not.	111
7.7	Daily Click-Through Rate (CTR): <i>baseline</i> vs. <i>ad quality</i>	119
8.1	Preference matrix.	135
8.2	Reading patterns.	140
8.3	Stability of articles.	143
8.4	Transitions between reading patterns considering all articles and only articles with a high stability.	145
9.1	Differences between user-based networks.	157
9.2	Differences between time-based networks.	158
9.3	Differences between traffic-based networks.	159
9.4	Differences between upstream-based networks.	161
9.5	The strength of the network effect generated at different correlation thresholds.	163
9.6	Network effect patterns generated at different correlation thresholds.	165
9.7	Network effect patterns.	166
9.8	Percentage of link types depending on site category.	168
10.1	Selecting the threshold to decide when two articles relate to the same story, and the threshold to decide when a story is niche.	179
10.2	Difference in story-focused reading and multi-provider reading between the actual dataset and the shuffled dataset.	181
10.3	Story-focused reading and popularity, number of articles, and coverage.	182
10.4	Distribution of reading and story-focused sessions over the users.	184
10.5	Comparison of story-focused and non-story-focused sessions at different session lengths.	187

10.6 Number of clicks on inline links depending on the number of inline links of an article. 194

10.7 Popularity and performance of inline links depending on their position in the text. 196

10.8 Session activity (dwell time) and loyalty (absence time) of users depending on the type of provider session. 199

11.1 Distributions of the number of users per crowd. 209

11.2 Proportion of retweets during each crowd’s creation. 210

11.3 Distributions of number of tweets per day and different type of tweets. 212

11.4 Average number of followers and followees of users per crowd. Each data point corresponds to one crowd. 213

11.5 Depiction of our assignment of slices to tweets in the data. 214

11.6 Time granularity. 215

11.7 Correlation test. 216

11.8 Word clouds generated for the crowd on the AJE story “Central African rebels advance on capital”. 225

12.1 Key findings in each chapter of the dissertation. 238

List of Tables

4.1	Site engagement metrics capturing the popularity, activity, and loyalty of sites.	29
4.2	Kendall’s tau between engagement metrics.	32
4.3	Intersections of the patterns (cluster similarities).	41
5.1	Site categories and their subcategories, and percentages of sites in each (sub)category. For some subcategories a description is given as well.	48
5.2	Multitasking characteristics depending on the size of a session.	54
5.3	Session and site visit characteristics depending on the size of a session.	55
5.4	Metrics used to analyse multitasking behaviour within sessions.	59
5.5	Spearman’s rho between multitasking and activity metrics.	65
6.1	Network instances based on five countries in one continent.	78
6.2	Network-level metrics: Metrics used to analyse user engagement within a network.	84
6.3	Node-level metrics: Metrics used to analyse user engagement with a site in the network.	87
6.4	Spearman’s rho between network-level metrics.	89
6.5	Spearman’s rho between node-level metrics.	90
7.1	Changes in engagement for short and long ad clickers.	108
7.2	Prediction performance on models built on ads data from March 2014 and tested on April 2014.	115

7.3	Top-15 ranked features using random forest with two sets of features: C, and C-S.	116
7.4	Differences (%) in dwell time and bounce rate between <i>ad quality</i> and <i>baseline</i> ads, and <i>ad quality</i> and <i>baseline</i> users.	120
8.1	List of the 500 most popular articles in Wikipedia.	132
9.1	Network instances based on the US network.	154
9.2	Spearman’s rho between site metrics using the link and traffic network.	169
10.1	Sections in which it is more likely to find story-focused reading.	185
10.2	Statistics of story-focused sessions of different depth.	188
10.3	Percentage of upstream traffic sources as a function of the depth of a story-focused session.	191
10.4	Popularity and performance of different types of inline links.	195
11.1	Dataset characteristics: number of articles, users, and tweets.	208
11.2	Example of candidates found for two stories.	218
11.3	Evaluation of the discovery of related articles, in terms of area under the ROC curve and recall at 2/3 precisions (R@2/3).	222
11.4	Aggregated ranking of features by importance (most important first) across the three datasets.	223
11.5	Example of users for two news articles.	228
11.6	Distributions of the human-provided labels.	230
11.7	Evaluation of models for the <i>UserIsHuman</i> and <i>UserIsInterestedIn-Story</i> tasks.	233
A.1	Dwell Time prediction performance on models built on ads data from March 2014 and tested on April 2014.	277
A.2	Bounce rate prediction performance on models built on ads data from March 2014 and tested on April 2014.	278
A.3	CombHQ prediction performance on models built on ads data from March 2014 and tested on April 2014.	278
A.4	News providers under consideration, listed in alphabetical order.	279

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 being captivated by a website, and so being motivated to use it [195]. “In a world full of choices where the fleeting attention of the user becomes a prime resource, it is essential that [...] providers do not just design [websites] but that they design engaging experiences.” [13].

However, before we can design engaging websites, it is crucial that we are able to measure user engagement – as the physicist Sir William Thomson has already pointed out: “If you can measure it, you can improve it.” Indeed, measurements are essential for analysing whether a website is engaging or not, and they can be used to evaluate the impact of design changes on engagement. One way to measure user engagement is through online behaviour metrics aiming at assessing users’ depth of interaction with a website. Widely-used metrics include click-through rate, time spent on a site (dwell time), page views, return rates, and number of users. Although these metrics cannot explain why users engage with a website, they are extensively used by the web analytics community and Internet market research companies such as comScore as *proxy* for online user engagement.

This thesis starts with a study on online behaviour metrics, focusing particularly on their limitations and proposing new metrics that expose so far unconsidered aspects of user engagement. We then show, through case studies, how these metrics enhance our understanding of user engagement, and we study the relationships between website characteristics and user engagement, which provide ideas on how to design engaging websites.

1.1 Research Problems

In this section we motivate and define the research questions of the thesis. We pose broad and general research questions.

A common approach to measure user engagement is through online behaviour metrics [48; 122], such as the number of visitors, dwell time, and return rates. In this thesis, we refer to this type of metrics as *site engagement metrics* and to this type of user engagement as *site engagement*, as it is concerned with the engagement of users *at site level*.

Site engagement varies significantly and these variations do not necessarily entail that one site¹ is more engaging than another. Instead, these differences depend on various factors such as the goals of a site [219] and the audience it attracts. For example, users spend much less time on search sites than on social media sites, because for the former sites the main goal is to find information (search results) quickly and leave afterwards. In addition, even sites of the same type can differ in their engagement. For instance, Benvenuto [21] studied user online behaviour on social media sites and found that users spend much less time on LinkedIn than on MySpace. Finally, site engagement depends on the tasks users want to perform on a site [167]. For example, users visit Wikipedia in order to quickly check some facts, but also to edit Wikipedia articles, thus to be part of the Wikipedia community. Each type of task leads to a different type of site engagement.

Motivated by these observations, we define the following research questions:

[Q1] Research questions: *How does user engagement differ between sites and how can we measure and characterise such differences? What should be taken into consideration when measuring user engagement?*

One aspect of user engagement is “stickiness” [32], which is concerned with users’ depth of interaction with a site. Stickiness can be measured with metrics that capture the activity of users on a site, such as the time spent (dwell time) or the pages viewed during a visit. In this thesis, we refer to users’ activity on a site as their *browsing activity*.

It is important to be aware of the limitations of activity metrics given by the general online behaviour of users on the Web. For instance, it has been observed that users often access several sites during an online session [138; 264],

¹In this thesis, we use site and website interchangeably.

and many of these sites are visited more than once [192]. A user may visit several sites to perform a single task (*e.g.* to compare offers from different shopping sites) or to perform several totally unrelated task in parallel such as emailing, reading news, or contacting friends on a social network. In doing so, users switch between sites using, for instance, the back button or the browser tabs. In fact, tabs are particularly useful for this type of online behaviour, as users can leave several tabs (tasks) open and switch between them [70; 105].

We refer to this type of online behaviour as *online multitasking*, and we are interested in the following questions:

[Q2] Research questions: *Does online multitasking affect engagement metrics that capture users' browsing activity on a site? Is it possible to define metrics that characterise multitasking during online sessions and do these metrics provide new insights compared to standard engagement metrics?*

Engagement metrics were designed to measure engagement at site level, which implies that these metrics are not applicable for assessing engagement with more than one site. In the context of online multitasking, it might be interesting to understand how users engage with several sites when performing a single task and which implications this has for each site. For instance, even if users favour a certain news site for their daily news consumption, they may also visit other news or social media sites such as Twitter to find articles they are interested in [36; 139]. This might imply that users spend less time on their favourite news site, *i.e.* they are less engaged. However, this behaviour can also reveal opportunities for a site to increase its engagement. A news provider, for instance, can benefit from a successful presence on social media sites, as it will drive traffic to its own site [169].

It is important to measure engagement across sites even when the user is performing totally unrelated tasks on these sites. For instance, many large online providers (*e.g.* Amazon, Google, Yahoo) offer a variety of sites, ranging from news to shopping, and the aim of these providers is not only to engage users with each site, but with as many sites as possible; in other words, they want to increase the user traffic between sites.

Engagement metrics cannot measure such online behaviour, and how to adapt them to measure engagement in a network of sites is not obvious, as they do not account for the user traffic between sites. We therefore propose

a methodology for studying *inter-site engagement*, that is, user engagement within a network of sites. We address the following research questions:

[Q3] Research questions: *How can we measure user engagement with respect to a network of sites? How does this enhance our understanding of engagement?*

Another important objective of website owners is to design engaging websites and to continuously increase engagement. In this context, it is essential to understand which site characteristics affect engagement. Various studies exist that show the impact of, for instance, video quality [67], saliency (visual catchiness) of relevant information [172], or the navigation structure of a site [267] on user engagement.

The navigation structure of a site is especially interesting in the context of inter-site engagement. Although hyperlinks usually assist users when navigating through a website [118], they can also be used to direct users to other sites [279]. This implies that hyperlinks are probably a key element to influence inter-site engagement. In this thesis, we investigate this.

In addition, also the content provided by a site plays an important role in user engagement. For example, if a shopping site provides further product information (detailed description, reviews), it can encourage users to purchase something [249]. In addition, studies have shown that content personalisation increases engagement [84], and user satisfaction [161]. Indeed, the objective of recommender systems is to serve the most relevant item to a user, with the aim to keep him/her engaged – for the moment and in long-term [277]. Furthermore, engagement metrics can be used to increase the quality of content [48] by identifying pages on which users are bouncing of, or items with a low click-through rate.

In this dissertation, we provide further insights about the effect of content characteristics and hyperlinks on engagement, and we develop approaches that can have a positive impact on engagement. We focus particularly on our new dimensions of user engagement – online multitasking and inter-site engagement. We aim to answer the following research questions:

[Q4] Research questions: *How do the characteristics of a site (content and hyperlinks) affect user engagement? Can we use such dependencies to develop applications that have the potential to improve the engagement of users?*

1.2 Thesis Structure and Main Contributions

Now, we provide an overview about the structure of the thesis and highlight the main contributions of each chapter. The structure of the thesis is visualised in Figure 1.1. The fundamentals part and each chapter of the two application parts (top rectangles) provide answers related to the first three research questions ($Q1 - Q3$). The last research question ($Q4$) is studied in each chapter of the application parts (middle and bottom rectangles).

Part I: User Engagement

Chapter 2 presents existing research on user engagement, and it positions the thesis and its contributions in their context. In **Chapter 3**, a description of the methodology and the types of data used in this thesis is given.

Part II: Fundamentals

This part compares sites regarding their engagement characteristics with each other and studies the limitations of standard (site) engagement metrics. Taking into account these limitations, we define new metrics that account for online multitasking and inter-site engagement, and we show how they enhance our understanding of user engagement. Part of this work will be published as a book chapter:

[140] Mounia Lalmas and Janette Lehmann. “Models of User Engagement”. In H. L. O’Brien and M. Lalmas (Eds.), *Why Engagement Matters: Cross-disciplinary Perspectives and Innovations on User Engagement with Digital Media*. Springer, 2015, *in progress*.

Chapter 4: We study the diversity of user engagement through site engagement metrics that reflect the popularity, activity, and loyalty of a site. The effect of user type and temporal aspects on site engagement is analysed. We identify simple but intuitive patterns of engagement by clustering the sites based on their engagement characteristics. These patterns show how engagement differs across sites and highlights the important engagement characteristics of a site. Parts of this chapter were published in:

[148] Janette Lehmann, Mounia Lalmas, Elad Yom-Tov, and Georges Dupret. “Models of user engagement.” *International Conference on User Modeling, Adaptation, and Personalization (UMAP 2012)*, pp. 164-175, Montreal, Canada, July, 2012.

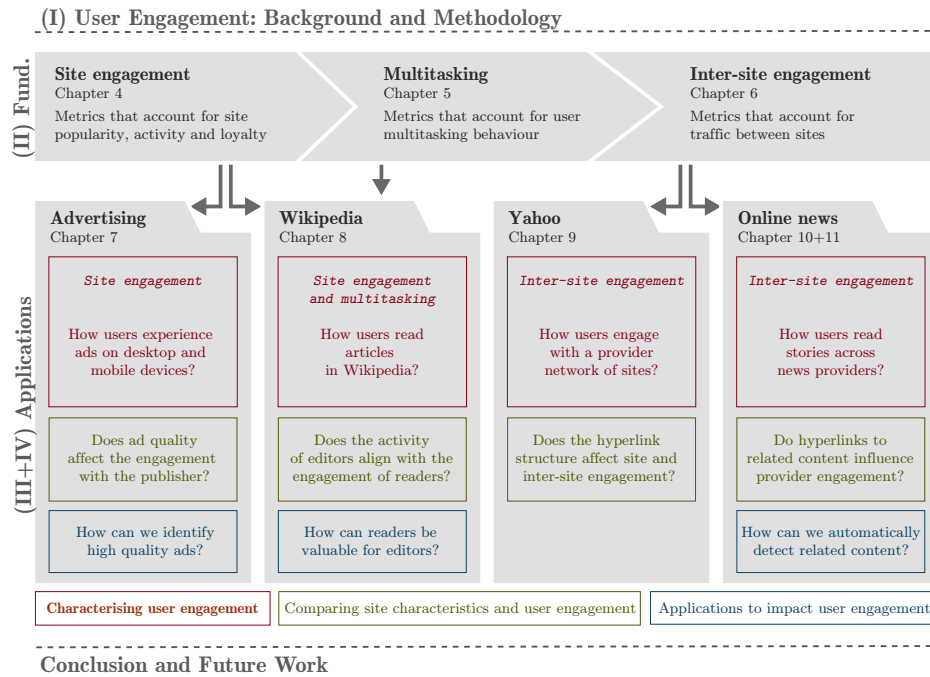


Figure 1.1: Structure of the thesis. Part I provides an overview about existing research and the methodology used. Part II introduces all engagement metrics used and forms the basis for the two application parts (part III and IV) which characterise users engagement in four case studies.

Chapter 5: This chapter describes the characteristics of online multitasking and shows how they affect site engagement metrics. We then define metrics that account for online multitasking, and we show that these metrics provide new insights about how users engage with a site. The following publication is based on the work described in this chapter:

[153] Janette Lehmann, Mounia Lalmas, Georges Dupret, and Ricardo Baeza-Yates. “Online multitasking and user engagement.” *ACM International Conference on Information and Knowledge Management (CIKM 2013)*, pp. 519-528, San Francisco, United States, October, 2013.

Chapter 6: This chapter studies the characteristics of inter-site engagement. We model sites (nodes) and user traffic (edges) between them as a network, and we propose a new set of metrics that account for the traffic between sites. We demonstrate the value of our approach using a large

provider network of sites offered by Yahoo. This chapter has been described in the following three forums:

[155] Janette Lehmann, Mounia Lalmas, and Ricardo Baeza-Yates. “Measuring Inter-Site Engagement.”. In V. Govindaraju, V. V. Raghavan, and C. R. Rao (Eds.), *Handbook of Statistics*, Elsevier, 2015.

[152] Janette Lehmann, Mounia Lalmas, Ricardo Baeza-Yates, and Elad Yom-Tov. “Networked User Engagement.”, *ACM Workshop on User engagement optimization at CIKM*, pp. 7-10, San Francisco, United States, October, 2013.

[151] Janette Lehmann, Mounia Lalmas, and Ricardo Baeza-Yates. “Temporal Variations in Networked User Engagement.”, *TNETS Satellite at European Conference on Complex Systems (ECCS)*, Barcelona, Spain, September, 2013.

Part III: Applications I: Site Engagement

We use site engagement and multitasking metrics for two case studies. The first case study analyses how users experience advertisements on websites on different devices and the second case study investigates how users engage with Wikipedia articles. This part of the dissertation also studies how the characteristics of ad landing pages and Wikipedia articles can affect engagement, and how these characteristics can be used to impact user engagement.

Chapter 7: In this chapter the relationship between engagement and advertising is analysed. The post-click experience on ads is measured through well-known engagement metrics: dwell time and bounce rate. The study analyses whether the device (desktop vs. mobile) upon which an ad is served has an impact on how users experience the ad, and whether a poor post-click experience can have a negative effect on the user engagement with the publisher’s site. Finally, we propose a method to identify high quality ads by analysing their landing pages and relating these to the ad post-click experience. The resulting prediction model can prioritise high quality ads. The work of this chapter is included in:

[141] Mounia Lalmas, Janette Lehmann, Guy Shaked, Fabrizio Silvestri, and Gabriele Tolomei. “Measuring Post-click User Experience with Mobile Native Advertising on Streams.”, *submitted for publication*.

Chapter 8: This chapter analyses how users consume content on Wikipedia: their reading preferences and behaviour. Site engagement and multitasking metrics are employed to study the reading behaviour on articles and how users access and re-access articles. The identified reading patterns are compared with the characteristics of the articles such as their text length and quality. We discuss how this information can be used by Wikipedia editors in their editing tasks to further improve articles and hence readers engagement. Parts of this work were published in and presented at:

[156] Janette Lehmann, Claudia Müller-Birn, David Laniado, Mounia Lalmas, and Andreas Kaltenbrunner. “Reader preferences and behavior on Wikipedia.”, *ACM International Conference on Hypertext and Social Media (HT 2014)*, pp. 88-97, Santiago, Chile, September, 2014, Ted Nelson Newcomer Paper Award.


[157] Janette Lehmann, Claudia Müller-Birn, David Laniado, Mounia Lalmas, and Andreas Kaltenbrunner. “What and how users read: Transforming reading behavior into valuable feedback for the Wikipedia community.”, *Presentation at Wikimania*, London, UK, August, 2014.

Part IV: Applications II: Inter-Site Engagement

The last part of the thesis analyses inter-site engagement within the Yahoo network of sites and during online news consumption using the metrics defined in Chapter 6 and various other methods. In addition, we study the impact of the hyperlink structure of sites on inter-site engagement, and we show how the concept of inter-site engagement can be used to develop an application that has the potential to improve the engagement with a site.

Chapter 9: This chapter is about inter-site engagement within the Yahoo network of sites. We observed in Chapter 4 that site engagement depends on factors such as the user type and periodic variations over time. This work analyses whether the same can be observed with respect to inter-site engagement. The study also compares the hyperlink with the traffic network and shows that hyperlinks can influence user online behaviour in the network. This chapter will be published as a book chapter:

[155] Janette Lehmann, Mounia Lalmas, and Ricardo Baeza-Yates. “Measuring Inter-Site Engagement.”. In V. Govindaraju, V. V. Raghavan, and C. R. Rao (Eds.), *Handbook of Statistics*, Elsevier, 2015.



Chapter 10: Our aim is to analyse inter-site engagement with respect to online news consumption, that is, how users read news from diverse news sites and other sources. We focus on a specific phenomenon: users reading several articles related to a particular news development, which we call story-focused reading. The study characterises story-focused reading and shows the differences compared to non-story-focused reading. We analyse how news sites promote story-focused reading, by looking at how they link their articles to other related content either published by them or other sources, and the effect of it on user engagement. The work of this chapter is included in:

[154] Janette Lehmann, Carlos Castillo, Mounia Lalmas, and Ricardo Baeza-Yates. “Story-Focused Reading in Online News.”, *submitted for publication*.

Chapter 11: The results of the previous chapter show that providing related content on an article page promotes story-focused reading and as a result keeps users engaged. This chapter shows how inter-site engagement, more precisely, the fact that users also visit social media sites for their daily news consumption, can be used to identify articles and other content related to a news story. We are interested in Twitter as a medium to help journalists and news editors in rapidly acquiring related content and information about the articles they publish. This information can complement or extend the articles they publish which in return can have a positive impact on engagement. The work described in this chapter was published in:

[149] Janette Lehmann, Carlos Castillo, Mounia Lalmas, and Ethan Zuckerman. “Transient News Crowds in Social Media.” *International AAAI Conference on Weblogs and Social Media (ICWSM 2013)*, Boston, USA, July, 2013.

[150] Janette Lehmann, Carlos Castillo, Mounia Lalmas, and Ethan Zuckerman. “Finding News Curators in Twitter.” *ACM International Conference on World Wide Web Companion (WWW 2013 Companion)*, 863-870, Rio de Janeiro, Brazil, May, 2013.

Chapter 12: Conclusions and Future Work

The dissertation ends with conclusions and thoughts for future work. The main results of the thesis are presented, and answers for each research question are given. Finally, we discuss various directions for future research.



PART I

User Engagement

This part discusses existing research on user engagement, and it provides a description of the datasets and the methodology commonly used in this thesis.



Background

This chapter provides an overview of existing research about user engagement. Work on user engagement can be divided into the following three areas: Defining user engagement, measuring user engagement, and improving user engagement. In the following sections we discuss the three areas, and position the thesis in their context.

2.1 Definition of User Engagement

Although until today no clear definition of “user engagement” exists, researchers agree that engagement emphasises the positive aspects of the interaction between a human and a machine.

Existing research describes engagement as the cognitive and affective involvement of users in an interaction. For instance, user engagement has been connected with the flow theory [39; 268], where flow is defined as the mental state in which a user is fully immersed [54]. Laurel [143] considers, beside the cognitive state, also the affective involvement and describes engagement as the state of mind in which a user is enjoying an interaction. Further works [68; 215; 268] also emphasise that engagement is accompanied by a pleasurable experience. In addition, engagement can be defined by the behavioural involvement of users. For instance, Rodden *et al.* [219] defines engagement as “user’s level of involvement with a product” and describes that it can be analysed through the observed behaviour (*e.g.* number of visits per week) of a user towards a product. The online industry also

defines engagement as “an estimate of the degree and depth of visitor interaction on the site against a clearly defined set of goals” [206]. In this context, the interaction design of a product is playing an important role in engagement [110; 222].

Recently, O’Brien *et al.* [195] combined definitions from various disciplines under the aspect of flow, aesthetic, play, and information interaction theory and came to the conclusion that user engagement can be characterised by the following attributes: aesthetic and sensory appeal, affect, awareness, challenge, feedback, interactivity, interest, motivation, novelty, and perceived control and time. These attributes were further extended by Attfield *et al.* [13] to include user context, reputation, trust, and expectation. Inspired by this and other definitions, Attfield *et al.* [13] and Lalmas *et al.* [142] define user engagement as:

“User engagement is the emotional, cognitive
and behavioural connection that exists, at any point in time
and over time, between a user and a resource.”

The thesis studies one aspect of this definition: We analyse the “behavioural connection” between users and a website (resource) using engagement metrics that capture users’ online behaviour on that site. These metrics characterise the online behaviour of users when visiting a site (“at any point in time”) and across visits (“over time”). Since this thesis is about online behaviour metrics and their limitations, we do not study the emotional and cognitive connection of engagement, since other types of measures (*e.g.* questionnaires, eye tracking) would be required to capture these aspects of engagement.

2.2 Measurements

Approaches to measure user engagement can be divided into three main groups [142]: self-reported, physiological, and online behaviour methods. This thesis focused on the latter group – online behaviour measures. For completeness, we provide a brief overview of the other methods, followed by a detailed description of online behaviour measures.

Self-reported Methods

In this group of methods, questionnaires and interviews (e.g. [196; 228]) are used to elicit user engagement attributes and to evaluate engagement immediately before, after, or while interacting with a website. The methods can be carried out within a lab setting or through online mechanisms (including crowd-sourcing). Notable works are provided by O'Brien *et al.* (e.g. [196; 193]). They developed a questionnaire for the evaluation of user engagement. The so-called user engagement scale has been used to study various applications, such as online shopping [193], the social media site Facebook [16], and web search [197]. The results also demonstrate that, depending on the website under consideration, different dimensions of engagement are important.

Although self-reported approaches are most suitable to evaluate users' emotions and thoughts about a website, these methods have known drawbacks, such as the reliance on user subjectivity [233], and the sensitivity to the *Mere-exposure effect* [250].

Physiological Methods

The second approach uses observational methods (e.g. facial expression and speech analysis) and neuro-physiological measures utilising tools such as eye tracking [74], heart rate monitoring, and mouse tracking [106]. Eye and mouse tracking are particularly useful to evaluate users' cognitive engagement in the web context, since these methods enable analyses of user attention on web pages [187]. Several studies (e.g. [106; 188]) have found correlations between mouse and eye movements on a web page. Moreover, Huang *et al.* [106] showed that cursor movements can be used to estimate the relevance of search result, even in the absence of clicks on the result page. This may suggest that even without any click activity of users (which is required for online behaviour methods), it is possible to examine the engagement of users to a web page. More interestingly, further works have demonstrated that mouse and eye movements can be used to identify frustrating or distracting experiences [187], and also to measure user engagement [9; 11].

However, physiological measures, although objective, have some limitations. Most of them are designed as lab studies, and hence the studies involve only a small number of participants and the laboratory setting might influence the user behaviour, apart for those who can be carried out with a software on the computer to track users' behaviour with a site (e.g. mouse tracking).

Online Behaviour Methods

In the online industry, engagement is measured through online behaviour metrics aiming at assessing users' depth of interaction with a site. Several reports contain studies on existing online behaviour metrics and their usage (*e.g.* [98; 206]). More recently, researchers have started to use these metrics to measure user engagement with, for instance, Q&A websites [72], online videos [5; 67], and content provided on a front page [277]. Only online behavioural metrics, referred to as engagement metrics, are scalable to millions of users, and are commonly employed as a *proxy* for user engagement with websites. Indeed, the fact that, for example, two million users choose to access a site daily is an indication of a high engagement with that site. Major websites are compared on this basis.

Early engagement metrics assessed the overall traffic on a site [269], such as the number of unique users, the number of visits, and the click-through rate. Although this group of measures is accounting for site *popularity*, they are still useful in the context of user engagement to measure how many users engage with the site.

The second type of metrics characterises the *browsing activity* of users when visiting a site. Widely used metrics are the number of page views and the time spend (dwell time) on a site. It should be noticed that the increasing use of JavaScript and XML (Ajax), and the integration of videos and slideshows diminishing the accuracy of the page view metric [73]. This makes dwell time the most important measure for visit activity nowadays. Several studies (*e.g.* [46; 278]) have demonstrated that dwell time on a site is a good indicator of engagement. Other works use dwell time to determine the relevance and quality of search results [167], advertisements [229], videos [5; 180], and other items [277; 278].

The last group of metrics captures the *loyalty* (retention) of users to a site. Typical metrics are return rate [84], absence time [72], and the number of visits of a user over a defined time period [32]. User loyalty refers to the endurability of engagement [142; 215], that is, users remember engaging experiences and want to repeat them. It has been shown [32] that there are interdependencies between loyalty and users' browsing activities during visits: users that visit a site frequently become familiar with that site which leads to less page views per visit. However, the dwell time remains the same.

The three types of metrics have been widely used to analyse user engagement with a site, and the results suggest that engagement depends on the site at hand. Whereas for many sites a high dwell time is an indicator of engagement [208; 279], search engines and digital libraries [126] desire a short dwell time and few clicks on result pages, since it implies that users quickly found the information they are searching. On the other hand, a long dwell time on a search result page is an important predictor for its relevance and interestingness [82]. Benevenuto *et al.* [21] characterised online behaviour on social media sites and found that there are significant differences; users spend much more time on Hi5 and MySpace than on LinkedIn. It is apparent that the “stickiness” of the content [32] on MySpace and Hi5 is responsible for this difference. These social media sites provide videos, photos, blog entries, *etc.*, whereas LinkedIn focuses on job offers and company pages. User online behaviour has been also studied in the context of online shopping. It has been shown that loyal users are more likely to purchase items [184], and that profitable shopping sites are more engaging with respect to the total number users, the dwell time on the site, and other metrics [89].

This thesis extends the study of user engagement through online behaviour metrics. We reveal the limitations of existing engagement metrics, and develop new metrics that enable us to study user engagement in a wider context (Chapter 5 and Chapter 6). We explore how sites differ in their engagement and how these differences can be characterised. Finally, we apply these metrics in four case studies to provide new insights about how users engage with sites (Chapters 7 to 10).

2.3 Factors of Influence

We discuss which factors influence engagement with a site. Understanding which factors influence engagement is essential to make informed decisions for improving engagement. In general, these factors can be categorised into provider context, user context, and website design.

The *provider context* is concerned with the “reputation, trust and expectation” [13] a user has about a website provider. Several studies have shown (*e.g.* [88; 135]) that trust is an important determinant of website success, especially in the area of e-commerce. If users do not trust a provider, they do not purchase. In this context, reputation [135] and brand [234] can have a positive effect on trust.

The *user context* describes how “user’s motivation, incentives, and benefits” [13] affect engagement. Attfield *et al.* [13] mentioned, for instance, that the same experience on another day or using another device can be different [145]. In addition, it also depends on current trends, cultural differences [86; 96], demographical factors (especially age and gender) [56], and how experienced the user is with the Web [191].

Finally, the *website design* has an impact on user engagement. Especially engagement attributes [197] such as aesthetic appeal, richness and control, and (content) novelty can be affected by the design of a website. Researchers investigated in what makes a website engaging [147], and showed that website design can influence trust, satisfaction, and loyalty [55; 183; 191], and users’ intent to revisit a site [86; 97]. Users can decide within the first 50 milliseconds on a site whether to engage with it or not [163], and this first impression (whether positive or negative) can impact a user’s attitude towards a website for a long time (halo effect) [58; 250]. Motivated by this, several guidelines have been developed that describe principles for the design of engaging experiences in general [245] and in the context of social media [208].

Two important perspectives are the navigational and informational design of a website. We describe next how we study these perspectives in the thesis.

Navigation and Hyperlinks

The navigational design refers to the navigation structure of a site formed by the hyperlinks on the pages. A number of works provide evidence that an unusable navigation can lead to users getting frustrated and then leaving the site [116; 267]. Therefore, efforts have been put in defining measures for website navigability [75], and in developing rules to create an intuitive and consistent navigation [118; 136].

Although the navigational design focuses mainly on the navigation menus, in a broader context, all hyperlinks within a website are important. Hyperlinks embedded in the text of a page can highlight the important content of the text [78; 79], and also the hyperlinks on the front page should be selected carefully, as these links drive traffic to other pages on the site [248]. Moreover, hyperlinks can be used to keep users on a site to increase user engagement [255; 279]. To summarise, hyperlinks are a simple yet powerful tool to direct users through the Web and through a website; they define the “marketplace of attention” [254], that is, users’ attention on the sites or pages where hyperlinks are pointing to [202].

We also study the hyperlink structure of websites, but focus on how hyperlinks can be used to influence site and inter-site engagement. In Chapter 9 we investigate how the hyperlink structure of the Yahoo network of sites can encourage users to visit other sites in the network, that is, to positively impact inter-site engagement. Chapter 10 explores how hyperlinks embedded in the text of a news article page affect site engagement – in terms of dwell time during a reading session (activity) and the time it takes until users return to the news site (loyalty).

Information and Recommendation

The information on a site, for instance its novelty, accuracy, and completeness, can influence user engagement. Indeed, an often discussed problem is the quality of user-generated content in general [60; 185] and in Wikipedia [2]. Studies focusing on user shopping behaviour have shown that providing information about a product (*e.g.* description, reviews) encourages users to purchase it [249]. It reduces users' uncertainty about the product, and enables him/her to make an informed purchasing decision. Today, in fact, many users go to Amazon to do product research (*e.g.* reading the reviews), even when they want to purchase the product somewhere else [208].

Similar observations can be made in the context of online news reading. To satisfy users' information needs, news sites provide additional content related to a story by linking to other news articles or information sources [38; 174]. This allows users to learn about the background of a story, or to gather different opinions around it.

Recommender systems can be used to increase content quality by selecting well performing content in general (*e.g.* news article with a high click-through rate) or personalised to the user. This has been shown to increase user engagement [84; 277] and user satisfaction [161]. Content recommendations are also useful to engage new users by recommending, for instance, people and profile entries to users that just signed-up on a social media site [84], or by recommending news articles to new users depending on where they are coming from (*i.e.* using the referrer URL) [252].

Another aim of recommender systems is to increase revenue by selecting products that users might want to buy [263; 286], or advertisements on those users are likely to click on [108; 235]. In this context, it is important to balance between revenue and engagement, as, for instance, annoying ads might have a negative impact on engagement [33; 56; 91].

Our contribution to this research is manifold. We show that content quality, more precisely serving high quality ads and providing related content to a news story, has a positive impact on user engagement (Chapter 7 and Chapter 10, respectively). Motivated by this observation, we propose two recommenders, whereas the first one identifies high quality ads (Chapter 7) and the second one suggests related content to a news story (Chapter 11). In Chapter 8, we also analyse whether and how the characteristics of Wikipedia articles affect reader engagement, and we discuss how this information can be used by the community for their editorial work.

This chapter provided an overview of existing research about user engagement and adjacent fields. We started with defining user engagement and specifying which aspects of this definition are analysed in the thesis. We then discussed approaches to measure user engagement, focusing in particular on online behaviour metrics. Finally, we reviewed work about factors that can influence engagement, whereby the thesis is concerned with how content quality and hyperlinks affect site and inter-site engagement.

Methodology

This chapter describes the data sources and methodology commonly used in the thesis.

3.1 Interaction Data

All studies in this thesis are based on anonymised interaction data, also known as clickstream data. Interaction data consists of the activities a user has done while browsing through the Web. In our case, the recorded activities are page views, represented as log entries of the following form:

$$(BCookie, Timestamp, URL, ReferrerURL)$$

where *BCookie* is a browser cookie, that is, a unique ID for each user. The *Timestamp* indicates “when” the page was viewed, and the *URL* refers to the web address of that page. In case the user was following a hyperlink to arrive to the page, the preceding page view (*i.e.* *ReferrerURL*) is saved as well. Otherwise, no referrer URL is given, indicating that the user jumped directly to the page using, for instance, a bookmark. The time the user spends on the page (dwell time) is the duration of time between that page view and the next page view. As we are studying engagement at site level, we extracted from each page view the first level of the subdomain (*e.g.* wikipedia.org) that was visited. For larger portals (*e.g.* AOL, Google, MSN, Yahoo!) we considered the second level of the subdomain (*e.g.* mail.yahoo.com), since these sites provide numerous services (*e.g.* search, mail, news).

The top part of Figure 3.1 displays an example of interaction data of a user with the bcookie ID *bc0*. We can see, for instance, that the user visited the BBC website to view two news articles at timestamp *t2* and *t3*. The log entry at *t3* has a referrer URL, showing that the user clicked on a hyperlink in the first news article page to navigate to the second news article. The same applies to the Wikipedia articles visited at *t5* and *t6*. For all other log entries no referrer URL exists, indicating that no hyperlinks were used to navigate to the corresponding pages.

We consider two sources of interaction data, described next. The corresponding datasets are described (*e.g.* time period, size) in each chapter.

Yahoo toolbar. We collected client-side interaction data from a sample of users who gave their consent to provide browsing data through the toolbar of Yahoo. The interaction data are particularly valuable, as they contain the activity of users during their whole online sessions (*i.e.* also outside of Yahoo). Therefore, we are able (1) to study user online behaviour across sites, and (2) to define the dwell time for almost all page views (except in case of a session end). Server-site interaction data, on the other hand, only record the activity on a certain site, and the dwell time on the last page viewed during a site visit is always missing [121; 269]. The Yahoo toolbar data represents the main data source of the thesis. To ensure that no strong bias in the interaction data affects our results and their applicability, we compare in Chapter 8 and Chapter 10 other data sources with our data. The comparisons show that the insights gained in the thesis are not specific to Yahoo toolbar users.

Yahoo news stream. In Chapter 7 we study how native ads embedded in the news stream on Yahoo's front page affect user engagement. We also investigate whether users experience ads differently depending on the device they are using (desktop vs. mobile). For this study, it was necessary to use server-side interaction data of Yahoo containing all interactions within the news stream of Yahoo's front page, on desktop and mobile. The study is not restricted by the fact that the data are collected on the server-side, as it focuses only on the interactions with the news stream.

3.2 Online Sessions and Site Visits

The user activities, that is, all interactions performed by a user (*BCookie*), were split into *online sessions*, where a session is a sequence of pages visited by a user until he or she goes offline. Following [35], a user is said to have

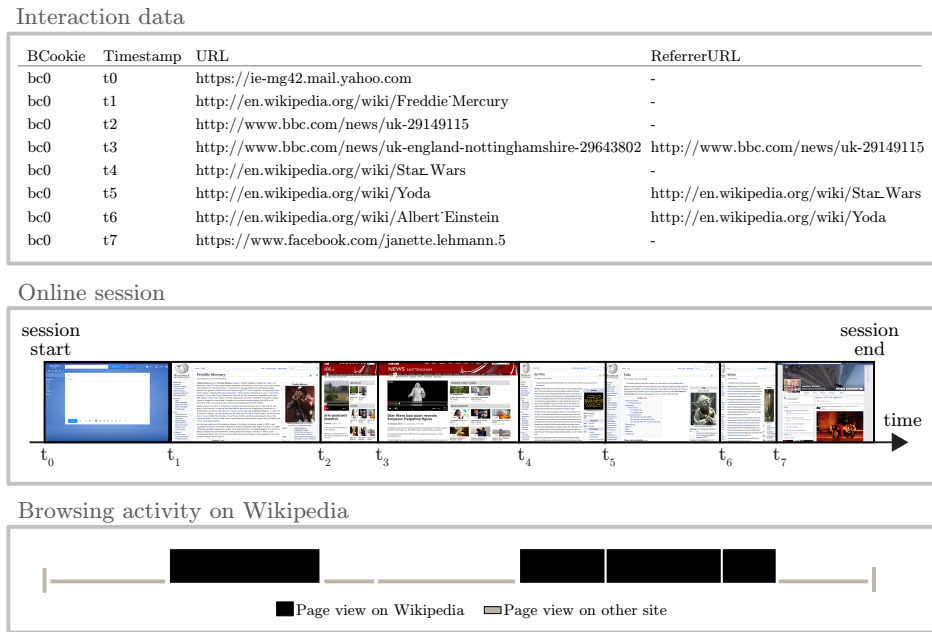


Figure 3.1: Example of interaction data: (Top) Log entries of the data. (Middle) Corresponding online session. (Bottom) Browsing activity on Wikipedia during that session.

gone offline – meaning that the session has ended – if more than 30 minutes have elapsed between two successive activities of that user. Consecutive page views to the same site are merged to form a *site visit*. The dwell time during a site visit is the time between the first page view of the visit and the subsequent page view after the visit. A site visit consists of the browsing activity of a user on a site, whereas the browsing activity informs us about the depth of engagement with that site. In addition, the visits over time assess the loyalty of the user.

In Figure 3.1 we show an example of an online session of a user (middle part), and the browsing activity on Wikipedia within that session (bottom part). We can see that Wikipedia was visited two times during that online session, whereas the visits consist of 1 and 3 page views, respectively. The corresponding dwell times during the visits are $(t_2 - t_1)$ and $(t_7 - t_4)$, respectively. In addition, the user visited further sites during that session, namely Yahoo mail, BBC, and Facebook.

3.3 Website Taxonomy

One aim of our work is to investigate whether sites of the same type (*e.g.* shopping sites) exhibit the same engagement characteristics, and whether the engagement characteristics differ among sites (*e.g.* shopping vs. news sites). Several approaches have been already developed to classify sites [138] or web activities [164] and the value of these taxonomies have been shown.

In this thesis, we adopt a similar taxonomy using three publicly available schemas: The Open Directory Project¹, the Yahoo directory², and Alexa’s ranking³ of the top sites per category. This resulted in site categories such as news, leisure, social media, search, and shopping. For some sites a manual annotation was necessary, and some categories needed to be added. For instance, in Chapter 6, we analyse inter-site engagement in the Yahoo network of sites. For the study we introduced the site category “provider” which relates to provider-related sites such as the Yahoo help or account site. In addition, the categorisation schema was adapted depending on the focus of the study. For example, in Chapter 10, it was necessary to consider much more news sites than in the other studies (1087 in total) as we analysed how users consume news on the Web. We introduce the underlying categorisation schema in each chapter.

Predominant site categories. In the following chapters of the thesis, sites are grouped using various criteria, such as the loyalty of users to the site, or the engagement characteristics over time. The predominant site categories per group are defined as follows: We define $p(c)$ as the probability that a site belongs to category c , and $p(c|g)$ as the probability that a site in group g belongs to category c . We then define the difference in the probability PD as follows:

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

This corresponds to the likelihood that a category occurs in a given group with respect to its likelihood that it occurs at all. A high value of PD (maximum is 100%) indicates that the site predominates the group, while a low value of PD (minimum is -100%) corresponds to site categories that are not important for the group.

¹<http://www.dmoz.org/>

²<http://dir.yahoo.com/>

³<http://www.alexa.com/topsites/category>

PART II

Fundamentals

In this part, we compare sites regarding their engagement characteristics with each other, and we define new metrics that account for online multitasking and inter-site engagement.



Site Engagement

This chapter provides a better understanding of how users engage with websites. We refer to this type of engagement as *site engagement*, as the metrics are used to measure engagement *at* site level. The aim of this chapter is to demonstrate the diversity of site engagement through the identification and the study of *patterns of site engagement*.

4.1 Introduction

Site engagement metrics are widely used to measure user engagement. These include, for example, number of unique users, click-through rates, page views, and time spent on a website. Although these metrics actually measure web usage, they are commonly employed as *proxy* for online user engagement: the higher and the more frequent the usage, the more engaged is the user.

However, engagement possesses different characteristics depending on the website; *e.g.* how users engage with a mail tool or a news portal is very different. While the former site will likely have many short visits during the day (checking new emails), the latter is probably characterised by longer visits on specific times of the day (*e.g.* morning, and lunch time). In other words, the web tasks users accomplished on a site influence the engagement characteristics of that site. Several approaches were developed to classify tasks and build taxonomies [124; 164; 170]. For instance, Kumar *et al.* [124] have grouped web tasks according to the type of site, for example, into categories such as social media, search and shopping.

To conclude, the type of website has an essential effect on engagement metrics. As a result, we should not speak of one main approach to measure site engagement, *e.g.* through one fixed set of metrics, because engagement depends on the website at hand. However, the same engagement metrics are typically used for all types of websites, ignoring the diversity of experiences. In addition, discussion on the “right” engagement metrics is still going on, without any consensus on which metrics to be used to measure which types of engagement.

To this end, we analysed a large number of online sites, of various types (ranging from news to e-commerce to leisure). We first show the diversity of site engagement for these sites. We also show how the audience (user types) and the temporal dynamics differ between sites. To identify *patterns of site engagement*, we clustered all sites using three criteria (dimensions) of engagement (general, user types, temporal aspects). Our results show that we can effectively derive patterns of engagement, for which we can associate characteristics of the type of engagement.

A review of existing research is included in Chapter 2. The chapter is organised as follows. Sections 4.2 and 4.3 describe the data and engagement metrics used. Section 4.4 demonstrates the diversity of site engagement. Section 4.5 presents the methodology adopted to identify patterns of site engagement and the outcomes. Section 4.6 looks at relationships between patterns, providing further insights about types of engagement. The chapter finishes with a discussion.

4.2 Dataset

We collected anonymised data during July 2011 from a sample of approximately 2M users who gave their consent to provide browsing data through the Yahoo toolbar. We restrict ourselves to 80 sites with at least 100 distinct users per month and within the US.

Site categories. Based on the website taxonomy of Section 3.3 we define the following seven site categories:

- 23% news sites [News]
- 18% leisure and social media sites [Leisure]
- 14% service sites (*e.g.* translators, search, mail) [Service]
- 14% support sites (*e.g.* product support, app download) [Support]

Table 4.1: Site engagement metrics capturing the popularity, activity, and loyalty of sites. $|D|$ refers to the number of days in the considered time frame.

Metric	Description	Engagement	
		Low	High
Popularity (for a given time frame) [POP]			
#Users	Number of distinct users.	0	∞
#Visits	Number of visits.	0	∞
#Clicks	Number of clicks (page views).	0	∞
Activity [ACT]			
PageViewsV	Avg. number of page views per visit.	0	∞
DwellTimeV	Avg. time per visit.	0	∞
Loyalty (for a given time frame) [LOY]			
ActiveDays	Number of days a user visited the site.	0	$ D $
ReturnRate	Number of times a user visited the site.	0	∞

- 14% settings sites (*e.g.* profile setting, personalisation) [Settings]
- 10% front pages and site maps [Front page]
- 7% shopping sites [Shopping]

The site categories enable us to analyse how the type of site influences engagement.

4.3 Site Engagement Metrics

The metrics used in this chapter are listed in Table 4.1. As our aim is to identify engagement patterns, we restrict ourselves to a small set of widely reported metrics. A preliminary analysis (see Section 2.2) revealed that there are three common types of metrics to study engagement, reflecting *popularity*, *activity*, and *loyalty*. Popularity metrics measure how much a site is used, *e.g.* total number of users. The higher the number, the more popular the corresponding site. How a site is used is measured with activity metrics, *e.g.* average number of clicks per visit across all users. Loyalty metrics are concerned with how often users return to a site. An example is the return rate, *i.e.* average number of times users visited a site.¹ Loyalty and popularity metrics depend on the considered time interval, *e.g.* number of weeks considered. A highly engaging site is one with a high number of visits

¹A user can return several times on a site during the same day, hence this metric is different from the number of active days.

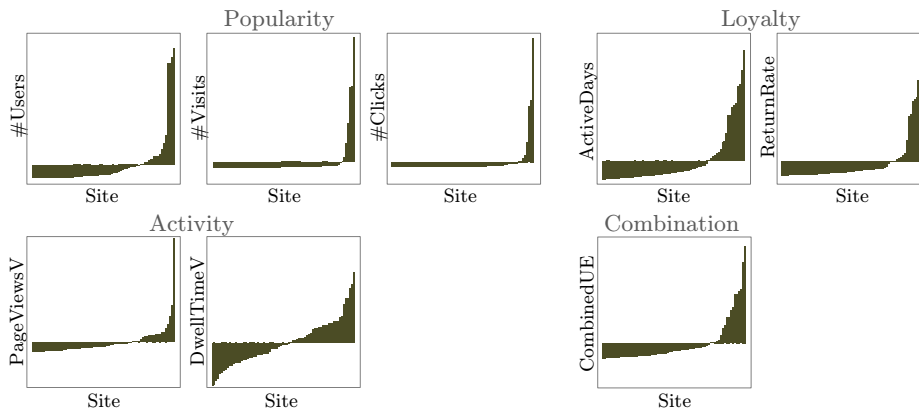


Figure 4.1: Normalised engagement values per site. The x-axes order sites by the metric value. The axes values are removed for confidentiality reasons.

(popular), where users spend lots of time (active), and return frequently (loyal). It is however the case, as demonstrated next, that not all sites, whether popular or not, have both active and loyal users, or vice versa. It does not mean that engagement with such sites is lower; it is simply different. Our conjecture is that site engagement depends on the site itself.

4.4 Diversity in Engagement

We show that engagement depends on the site under consideration and various other factors.

Sites. Figure 4.1 reports the engagement values for the seven metrics and the 80 sites under study. The x-axes in the plots represent the sites, ordered by increasing values for the corresponding metric. The values of a metric are normalised by the z-score, hence the plots show the extent to which the standard deviation of a metric value is above or below the mean. Finally, *CombinedUE* is the linear combination of *#Users*, *DwellTimeV*, and *ActiveDays*.

We can see that sites differ widely in terms of their engagement. The popularity of sites is not equally distributed over the sites. Some sites are very popular (*e.g.* news sites) whereas many others are visited by small groups of users (*e.g.* support sites). Loyalty per site is also very skewed. We speculate that, for instance, news and some service sites (*e.g.* mail, search) have

many users returning to them much more regularly, than sites containing information of temporary interests, such as shopping, support, and settings sites. Whereas *PageViewsV* exhibits a similar distribution as the popularity and loyalty metrics, the distribution of *DwellTimeV* is different. We can see that around 50% of the sites have a dwell time per visit below average, and the other 50% of sites have a dwell time above average. This implies that although many sites have a low popularity and loyalty, users spend time on them. However, they do not necessarily view many pages (*PageViewsV*). Finally, using one metric combining the three types of metrics (*CombinedUE*) also shows that engagement varies across sites.

Metrics. To show that engagement metrics capture different aspects of site engagement, we calculate the pair-wise metric correlations using the Kendall tau (τ) coefficient. We only consider correlations that are statistically significant (p-value < 0.05). The results are reported in Table 4.2.

First, we observed that metrics of the same group describe similar engagement characteristics, whereas metrics of different groups characterise engagement in different ways. Indeed, we can report that the resulting average intra-group correlation is $\tau = 0.71$, *i.e.* metrics of the same groups mostly correlate; whereas the average inter-group correlation is $\tau = 0.21$, *i.e.* metrics from different groups correlate weakly or not at all. Interestingly, the same can be observed when only considering sites belonging to a certain category. For news sites, for instance, we have an intra-group correlation of $\tau = 0.73$, and an inter-group correlation of $\tau = 0.43$. This implies that not only the type of site is responsible for the differences, but also further aspects affect the engagement of a site.

The three popularity metrics show a similar engagement type for all sites, *i.e.* high number of users implies high number of visits ($\tau = 0.82$), and vice versa. For the loyalty metrics, users that have more active days also return regularly to the site ($\tau = 0.79$). The correlation between the two activity metrics is lower ($\tau = 0.33$). We argue that dwell time is a more accurate measure of browsing activity, because the concept of page view is not well defined by dynamic changes to a web page such as in Ajax. As a result, the metric does not correlate with other activity measures.

The correlations between metrics from different groups are lower. There are no correlations between the activity metrics, and the popularity or loyalty metrics. We even cannot report correlations between most metrics, as the p-value is above 0.05. High popularity does not entail high activity ($\tau =$

Table 4.2: Kendall’s tau between engagement metrics. Correlations with a $p\text{-value} \geq 0.05$ are not reported (-).

	[POP] #Users	[POP] #Visits	[POP] #Clicks	[ACT] PageViewsV	[ACT] DwellTimeV	[LOY] ActiveDays	[LOY] ReturnRate	CombinedUE
#Users [POP]		0.82	0.75	-	-	0.43	0.34	0.56
#Visits [POP]	0.82		0.85	-	-	0.60	0.52	0.73
#Clicks [POP]	0.75	0.85		0.16	0.18	0.59	0.51	0.71
PageViewsV [ACT]	-	-	0.16		0.33	-	-	-
DwellTimeV [ACT]	-	-	0.18	0.33		-	-	0.19
ActiveDays [LOY]	0.43	0.60	0.59	-	-		0.79	0.81
ReturnRate [LOY]	0.34	0.52	0.51	-	-	0.79		0.69
CombinedUE	0.56	0.73	0.71	-	0.19	0.81	0.69	

0.09). Many sites have many users spending little time on them; *e.g.* a search site is one where users come, submit a query, get the result, and, if satisfied, leave the site. This results in a low dwell time even though user expectations were entirely met. The same argument holds for Q&A or weather sites. What matters for such sites is their popularity. The highest correlations can be observed between the popularity and loyalty metrics (*e.g.* $\tau(\#Visits, ReturnRate) = 0.52$), indicating that popular sites are those to which users return regularly. However, the correlation is only moderate.

We look now at the correlations between the metric *MetricCombi* and the other metrics. Interestingly, the metric correlates strongly with the popularity and loyalty metrics, but there is no correlation to the activity metrics. Combining a popularity and loyalty metric might represent the two engagement dimensions in one metric. However, the activity on a site is not captured by this metric.

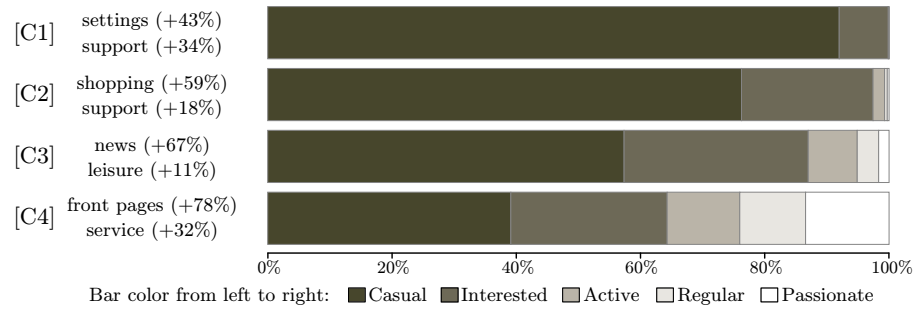


Figure 4.2: Site clusters and user groups (Casual, Interested, Active, Regular, Passionate). (Left) Cluster ID and for each cluster its predominant site categories and their probability differences in brackets.

Users. Studies have shown that users may arrive on a site by accident or through exploration [18], and simply never return. Other users may visit a site once a month, for example, a credit card site to check their balance. On the other hand, sites such as mail may be accessed by many users on a daily basis. We thus look at how active users are on a site within a month. The number of days a user visited a site over a month is used for this purpose. We create five types of user groups:²

- Casual: 1 day
- Interested: 2-4 days
- Active: 5-8 days
- Regular: 9-15 days
- Passionate: more than 16 active days

The percentage of the user groups for each site is calculated, and sites with similar percentage of user groups are clustered using k-means. Four clusters were detected and the cluster centers calculated. Figure 4.2 displays the four cluster centers, *i.e.* the percentage of user groups per cluster. On the

²The terminology and the proposed range of days is based on our experience in how engagement is studied in the online industry. For instance, a Passionate user is one that comes on average 4 days per week, thus leading to the figure of 16 days within a month.

left side the cluster ID and the predominant site categories per cluster are shown (see Section 3.3 for the definition of predominant site categories).

We observe that the percentage of Casual users is high for all sites. Cluster $C1$ has the highest percentage of Casual users; typical sites are support or settings sites (*i.e.* doing some profile settings). Cluster $C2$ includes sites related to specific tasks that are occasionally needed (*e.g.* purchasing an item); as such they are not visited regularly within a month. The third cluster ($C3$) includes sites related to leisure and news activities, which are used on a regular basis, albeit not daily. Finally, cluster $C4$ contains front pages and service sites (*e.g.* mail). For these sites, the percentage of Passionate users is higher than the percentage of Regular and Active users. The above indicates that the type of users, *e.g.* Casual vs. Passionate, matters when measuring engagement.

Time. The dynamics of online popularity has been studied in various contexts such as digital libraries [126] and Wikipedia [248]. Kulkarni *et al.* [137] identified, in the context of search queries, various patterns and classified them into spikiness, periodicity, overall trend, *etc.* In this section, we investigate how the temporal dynamics differ between sites.

Using the interaction data spanning from February 2011 to July 2011, we normalised the number of users per site ($\#Users$) by the total number of users who visited any of the sites on that day. The time series for each site was decomposed into three temporal components: periodic, trend and spikes, using local polynomial regression fitting [47]. To detect periodic patterns we calculated the correlation between the extracted periodic component and the residual between the original time series and the trend component. To detect spikes, the periodic component was removed from the time series and spikes were detected using a running median. Trending patterns were detected by comparing the extracted trending component with the residual between the original time series and the periodic component.

Figure 4.3 shows graphically the outcome for six sites (under examples), two for each temporal pattern, and the number of sites per pattern. Using the probability difference measure PD defined in Section 3.3, we selected for each temporal pattern two examples (high PD) and two counter-examples (low PD) of site categories (under categories). Finally, possible reasons for a periodic, spikiness, or trending pattern are given (under influence).

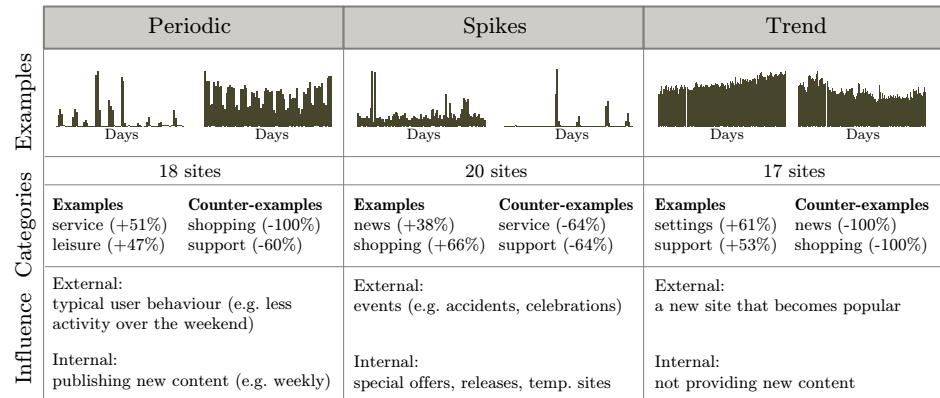


Figure 4.3: Engagement over time using $\#Users$ (February 2011 – July 2011): (Top) Examples of temporal behaviour types. (Middle) Examples and counter-examples of site categories. (Bottom) Reasons for the temporal behaviour type.

The engagement pattern can be influenced by external and internal factors. Service and leisure sites tend to be more “periodically used” than shopping sites. We observe that some sites are more popular during the weekend (*e.g.* leisure sites), whereas other sites are more used during the week (*e.g.* mail sites).

Shopping and news sites are characterised by spikes in popularity. Shopping sites are probably affected by internal factors, such as providing special offers for products over a short period. Access to news sites tends to be influenced by external factors (important news) or the frequency of publishing new information. Finally, some sites (*e.g.* settings and support) exhibit a trending behaviour, where the popularity is either increasing or decreasing. Sites that do not provide new content or that are newly released can have this type of temporal behaviour.

In summary, various reasons can cause the temporal dynamics in site popularity. With respect to our work, especially the periodic patterns are of interest, because they exhibit regular differences in engagement that are not necessarily caused by the site. A site that has lower popularity on weekdays does not have to be less engaging during the week, as it might be only a periodic behaviour. We therefore decided to account for time, more precisely the differences between weekdays and weekend, when measuring site engagement.

4.5 Patterns of Site Engagement

The previous section showed differences in site engagement. Now, we study these differences to identify patterns of site engagement. The base for all studies is a matrix containing data from the 80 sites under study. Each site is represented by seven engagement metrics. A metric can be further split into several dimensions based on user and time combinations (*e.g.* number of visits on weekdays versus weekends). The values of each metric are transformed into an ordinal scale to overcome scaling issues. We clustered the sites using the kernel k-means algorithm [65], with a Kendall *tau* rank correlation kernel [224]. The number of clusters are chosen based on the eigenvalue distribution of the kernel matrix. After clustering, each cluster center is computed using the average rank of cluster members (for each metric). To describe the centers (the patterns), we refer to the subset of metrics selected based on the correlations between them and the Kruskal-Wallis test with Bonferonni correction, which identifies values of metrics that are statistically significantly different for at least one cluster. For each pattern, we provide the predominant site category (see Section 3.3).

Three sets of patterns are presented, based on the seven engagement metrics (general), accounting for user groups (user-based), and capturing temporal aspects (time-based). Although all dimensions could be used together to derive one set of patterns (*e.g.* using dimension reduction to elicit the important characteristics of each pattern), generating the three sets separately provides clear and focused insights about engagement patterns.

4.5.1 General Patterns

First, we look at patterns of site engagement without accounting for user type or temporal aspect. We refer to them as *general patterns*. We use our seven metrics to generate six general patterns of site engagement, visualised in Figure 4.4. As the three popularity metrics exhibit the same effect, only *#Users* is reported. The same applies for the loyalty metrics, *i.e.* only *ActiveDays* is reported. The two activity metrics yield different behaviours, hence both are shown.

For pattern *G1* and pattern *G2*, low popularity is a main factor. A high number of page views but a low dwell time further characterise pattern *G1*. Support sites belong to this cluster, which are visited by users to, for instance, quickly download an application and then leave. In contrast, users

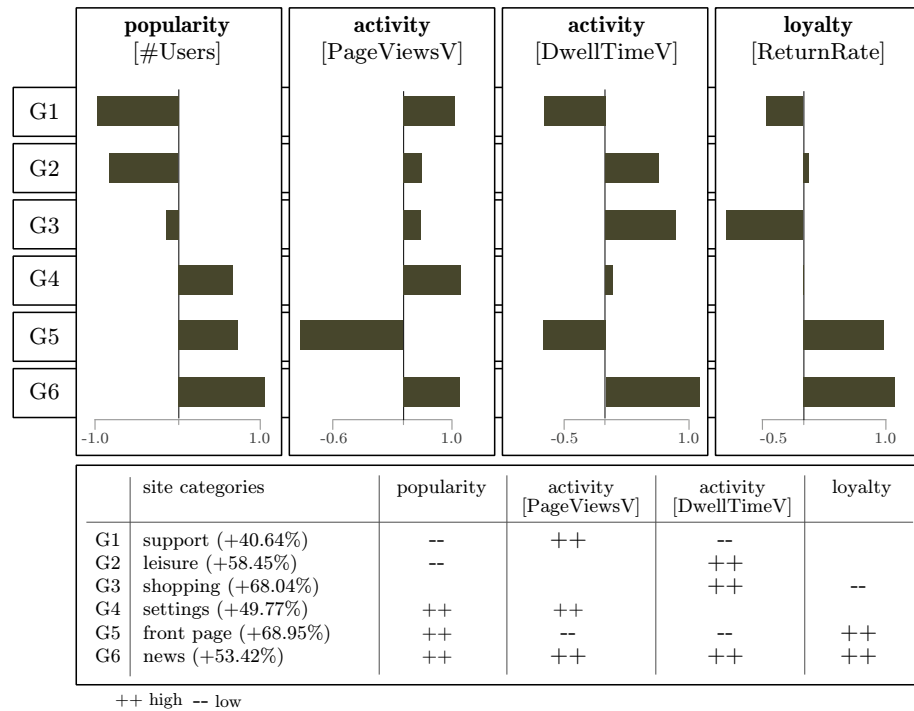


Figure 4.4: General patterns of engagement: (Top) Cluster centers representing the engagement characteristics. (Bottom) Pattern descriptions.

spend a lot of time on leisure sites (pattern *G2*) to play games, communicate with friends, *etc.*

Pattern *G3* describes sites where users spend time on, but with low loyalty. Shopping sites, which are not regularly (daily) accessed by users, follow this pattern. However, when accessed, users spend more time to select items they want to purchase. The main factor for pattern *G4* is a high popularity and a high number of clicks per visit. This pattern contains sites for personal settings (*e.g.* profile updating), hence, the main activity is to click.

Although pattern *G5* and pattern *G6* are both characterised by a high popularity and loyalty, the patterns differ with respect to their activity. Users spend little time and perform few page views on sites belonging to pattern *G5*. Front pages belong to this pattern; their role is to direct users to interesting content on other sites, and what matters is that users come regularly to them. In contrast, users view many pages and dwell long when consuming news (pattern *G6*).

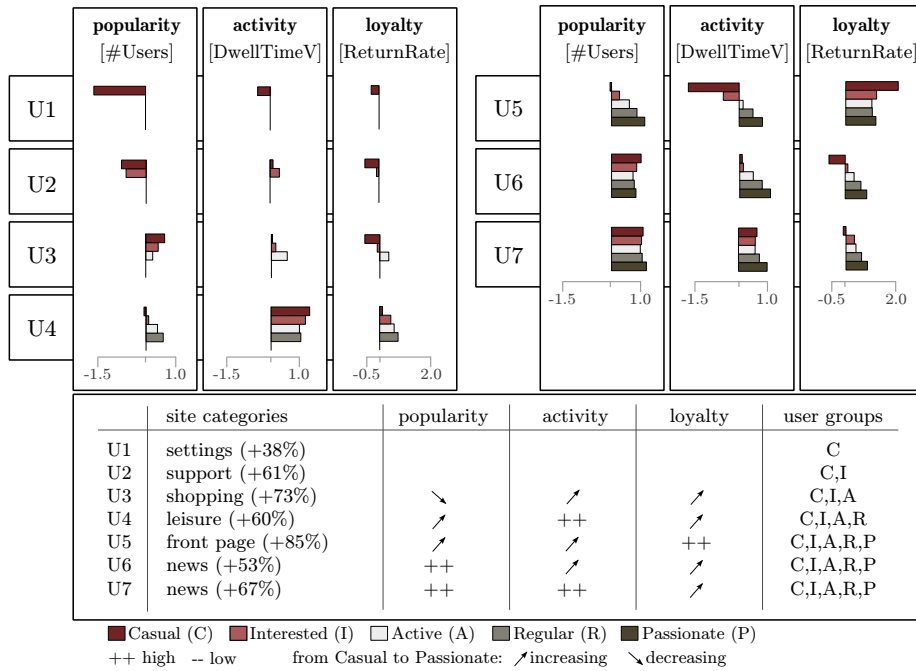


Figure 4.5: User-based patterns of engagement: (Top) Cluster centers representing the engagement characteristics. (Bottom) Pattern descriptions.

4.5.2 User-based Patterns

We investigate now patterns of site engagement that account for the five user groups elicited in Section 4.4. The seven metrics were split, each into five dimensions, one for each user group, *i.e.* Casuals to Passionate. This gives 35 engagement values per site. A site without a particular user group gets 0 values for all metrics for that group. We obtain seven *user-based patterns* (clusters), visualised in Figure 4.5. We only report the results for one metric of each group (*#Users*, *DwellTimeV* and *ReturnRate*), since these are sufficient for our discussion.

The first two patterns, pattern *U1* and pattern *U2*, are concerned with low engagement (popularity) of only Casual users, and Interested and Casual users, respectively. They correspond to sites on very particular interests or of a temporary nature (doing profile settings, downloading an application); as such popularity for these two groups of users is low compared to other patterns. Moreover, pattern *U1* indicates that, when on site, the activity of

Casual users is not negligible. By contrast, pattern *U2* highlights a higher activity of Interested users than Casual users.

Pattern *U3* caters for the engagement of Casual, Interested and Active users. Shopping sites belong to this pattern. Loyalty increases going from Casual to Active users, which makes sense as loyalty is used to determine the user groups. More interestingly is that activity augments the same way, whereas popularity decreases. This shows that it is less likely to have users that visit a shopping site frequently, but users that are loyal to a shopping site spend also more time on that site compared to less loyal users.

The next two patterns, pattern *U4* and pattern *U5*, exhibit the same increase in popularity from Casual to Regular (*U5*: Passionate) users. High activity across all user groups apart for Passionate and an increasing loyalty from Casual to Regular users is an important feature of pattern *U4*, which typically include leisure sites. High loyalty across all groups and an increase in activity from Casual to Passionate users further characterise pattern *U5*. Sites falling in this pattern include front pages.

Finally, pattern *U6* and pattern *U7* are characterised by high popularity across all user groups. Popular news sites belong to these patterns. Activity increases from Casual to Passionate users for pattern *U6*, whereas it is high across all user groups for pattern *U7*. It shows that for some news sites the time the user spends on the site increases with the user loyalty (*U6*), whereas for other news sites even Casual or Interested users spend a lot of time on the site (*U7*).

4.5.3 Time-based Patterns

We look now at patterns of site engagement that account for the temporal aspect. For simplicity, we consider two time dimensions, *weekdays* and *weekends*. Each site becomes associated with twelve metrics; six of our engagement metrics are split into these two time dimensions (*ActiveDays* is not used, as it has a different time span). To extract the differences in engagement on weekdays vs. weekends, we transformed the absolute engagement values into proportional ones, *e.g.* the proportional *ReturnRate* is $ReturnRate_{weekdays} / (ReturnRate_{weekdays} + ReturnRate_{weekend})$. The same methodology as that used for the other types of patterns was then applied. This led to the identification of five *time-based patterns* of engagement (clusters), shown in Figure 4.6.

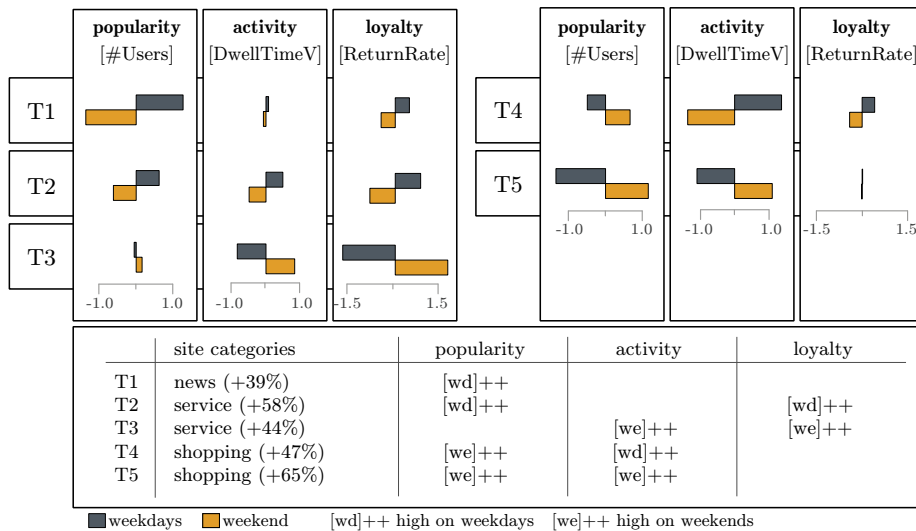


Figure 4.6: Time-based patterns of engagement: (Top) Cluster centers representing the engagement characteristics. (Bottom) Pattern descriptions.

We can see that pattern $T1$ and pattern $T2$ are similar as they both are characterised with higher popularity during weekdays. Sites related to news and service, respectively, follow these patterns. Pattern $T2$ is further characterised by higher loyalty during weekdays, because it contains sites used for work issues (*e.g.* search, mail). Pattern $T3$ also contains many service sites, which are characterised by a higher activity and loyalty during the weekend. Service sites following this pattern are about personal data management (*e.g.* calendar, address book).

Finally, pattern $T4$ and pattern $T5$ describe sites with higher popularity on weekends; activity is also higher on weekends for pattern $T5$, whereas it is higher on weekdays for pattern $T4$. Both patterns characterise sites related to shopping indicating that even sites of the same type can differ in their engagement.

4.6 Relationship between Patterns

First, we investigate whether sites belonging to one category (*e.g.* leisure, shopping) exhibit a certain engagement pattern or not. We can report that the sites of a category belong to on average 4.6 (median=5.0) different general patterns. A similar diversity can be observed for the user-based and

Table 4.3: Intersections of the patterns (cluster similarities). The value range is between 0 (minimal similarity) and 5.61 (maximal similarity).

	General patterns	User-based patterns	Time-based patterns
General patterns			
User-based patterns	3.48		4.12
Time-based patterns	4.12	4.25	

time-based patterns of engagement (4.4 and 3.9, respectively). This result is not surprising for categories that contain sites with very different objectives (*e.g.* mail and search are service sites). However, although shopping sites have the same objective (purchasing items), these sites follow three general patterns. The same can be observed for news sites, which follow six general patterns. This might show that one news site is more engaging than another, but it can also be a consequence of other factors such as the section of news they cover (*e.g.* some news sites focus on sport news), or the frequency in which new articles are published.

Finally, we checked whether the three groups of patterns describe different engagement aspects of the *same set* of sites or whether they are largely unrelated. We calculate the similarity between the three groups using the *Variance of Information*. The outcome is shown in Table 4.3 (5.61 is the maximal difference). We observe the highest (albeit low) similarity between the general and user-based patterns. The user- and time-based patterns differ mostly. Overall, all groups of patterns are independent, *i.e.* they characterise different if not orthogonal aspects of site engagement, even though the matrices used to generate them are related.

We do not show here all the relationships between each pattern of each group. Instead, we discuss two cases. For pattern *G3*, a general pattern characterising many shopping sites (high activity and low loyalty), 22% of its sites belong to the time-based pattern *T5* (also characterising mainly shopping sites with a high popularity and activity on the weekend), but further 22% follow the time-based pattern *T1* (high popularity on weekdays).

We now look at the user-based pattern *U4*, which characterises many leisure sites with high activity in all user groups and an increasing popularity and loyalty from Casual to Regular users. Sites following this pattern are split into three time-based patterns: pattern *T1* (33%) (high popularity on weekdays), pattern *T3* (45%) (high activity and loyalty on weekends), and pattern *T5* (22%) (high popularity and activity on weekends). This comparison

provides different angles into engagement, allowing to zoom into particular areas of interests, *e.g.* further differentiating the “high loyal” users associated with pattern $U4$ into weekdays vs. weekends.

4.7 Discussion

Our aim was to identify patterns of site engagement. We analysed a large sample of user interaction data on 80 online sites. We characterised site engagement in terms of three families of commonly adopted metrics that reflect different aspects of engagement: popularity, activity, and loyalty. We could show that the three types of metrics reflect different aspects of engagement, but metrics belonging to the same type characterise engagement in the same way. This is not the case for the two activity metrics, $PageViewsV$ and $DwellTimeV$, where we argue that the accuracy of the page view metric is limited [122]. It is possible that the page view metric is affected by the structure of the site. For instance, some sites split their content over several pages to increase the number of page views, and other sites integrate videos and slideshows on their pages, which keeps users longer on one page. In addition, the page view metric does not consider the dynamic changes on a web page using Ajax [73]. However, it does not imply that the page view metrics is useless, but that it can only be used when the structure and content of a site and its effects are known.

Using simple approaches (*e.g.* k-means clustering), we generated three groups of patterns of site engagement: general, user-based and time-based. This provided us different but complementary insights about user engagement and its diversity. We could see that sites differ widely in terms of their engagement. It is, however, not the case, as demonstrated, that user engagement on some sites is lower; it is simply different. Our conjecture is that user engagement depends on the site itself.

We want to note that this also applies for sites of the same type. For instance, following the web analytic company Alexa, by the time of this writing, Forbes is more popular than the Wall Street Journal (global rank is 148 and 264, respectively). However, users spend more time per day on wsj.com than on forbes.com (4:15min and 2:49min, respectively). We argue that these differences are caused by the fact that users need to register and pay for articles of the Wall Street Journal. Hence the site attracts less users, but users that visit the site are deeply engaged to it.

Looking at the general patterns, we could see that some sites are very popular (*e.g.* news sites) whereas others are visited by small groups of users (*e.g.* support sites). Visit activity also depends on the sites, *e.g.* search sites (as part of service sites) and front pages tend to have a much shorter dwell time than sites related to news and shopping. However, shopping sites are not visited frequently from a user, because shopping is not a regular (daily) activity. In contrast, many users return regularly to news sites and front pages. Loyalty is also influenced by the frequency in which new content is published (*e.g.* some sites produce new content once per week).

The user- and time-based patterns provided us with further insights. We could see, for instance, that the loyalty and activity of some news sites depend on each other; loyal users spend more time on the news site. However, for other news sites, even users that are not loyal to the site dwell long when visiting it. It also depends on the site whether users engage with it more during the week (*e.g.* news sites) or on the weekend (*e.g.* shopping sites). However, although more users visit news sites during the week (higher popularity), users that visit the sites on the weekend consume the same amount of news than during the week (similar activity).

Although each pattern describes the main characteristics of a certain type of site, we also observed that sites of the same type do not necessarily belong to the same pattern(s) of engagement. Differences in the type of content and the structure of the site might cause this diversity. This diversity can also point to sites that are not engaging because they are following the “wrong” pattern. Site providers should think carefully about their goals and which aspects of engagement are important for them, that is, to which engagement pattern their site should belong.

Limitations. Our work comes with certain limitations. The results might be influenced by the fact that we only considered a small sample of sites resulting also in a very simple site categorisation. Especially service sites differ significantly – with respect to the service they provide and their engagement. We therefore decided to extend our dataset and to use a more fine-grained categorisation schema in the following chapter.

We did not extract site features (related to site content and structure) and compared them with the engagement of sites. Therefore it is still difficult to explain why sites of the same type belong to different engagement patterns. Motivated by this, we decided to compare site features and engagement in

the application part of the thesis. This comparison also provides insights about how site features can be used to influence engagement.

Finally, we did not consider other types of metrics, such as sociometrics or metrics that relate to the activity on pages (*e.g.* number of comments on article page). These metrics might further enhance our understanding of user engagement. It is possible, for instance, that some sites have a highly engaged audience on Twitter, but these audience engage much less on the site itself. However, it is difficult to relate site and socio engagement, as some sites do not have a social media presence, and the focus of this work was to study how engagement differs between many different types of sites.

Online Multitasking

Users often access and re-access more than one site during an online session, effectively engaging in *online multitasking*. In this chapter, we study the effect of multitasking on site engagement, and we define new metrics that characterise such behaviour.

5.1 Introduction

When users are performing a task on the Web (*e.g.* planning a holiday), they may visit several sites (*e.g.* to compare offers from different travel sites, read reviews) within the session but also over several sessions before completing the task (*e.g.* to check offers over several days). In this chapter, we are concerned with users accessing several sites within an online session, *e.g.* emailing, reading news, accessing a social network. A user may access several sites to perform a main task or he or she may actually perform several totally unrelated tasks in parallel (*e.g.* responding to an email while reading news). This is very comparable to our general activities on desktop devices [200] and also to our daily life [223], where we often handle several tasks in parallel and we switch between them. This phenomenon can also be observed on the Web. We refer to both cases as *online multitasking*. In this chapter, we do not distinguish between the two cases; our focus is the effect of accessing and re-accessing several sites within an online session on the online behaviour of users. The metrics proposed in this chapter, however, provide insights about the type of multitasking.

Within a multitasking session, users can navigate between sites in several ways: *hyperlinking* (clicking on a link), *teleporting* (jumping to a page using

bookmarks or typing an URL), or *backpaging* (using the back button on the browser, or switching between open tabs or windows) [192; 283]. Backpaging implies that the navigation between sites is not linear, hence should not be modelled as standard linear click-streams. Referrer trees [177] were defined to model non-linear navigation. However, because engagement metrics are typically defined based on streams, and not trees, we “re-linearise” referrer trees into *tree-streams* and incorporate information about how users switch between sites. Tree-streams – as opposed to the commonly used click-streams – present a more accurate picture of the online behaviour of a user.

One aspect of site engagement is “stickiness”, which is concerned with users’ depth of interaction with a website. Typical metrics that characterise the stickiness of a site are the number of page views and the dwell time during a visit on that site. What we understand by multitasking has naturally consequences on how these metrics are calculated and on the conclusions we can draw.

We perform an extensive study of online multitasking and its effect on these two metrics. We examine how often a site is visited within a session, how many sites are visited in the same session, the type of sites being considered (to capture the effect of the task, *e.g.* reading news vs. doing emails), how users switch between sites, and whether any of these influence the assessment of site engagement. Our study is based on the interaction data of 2.5M users across 760 sites encompassing diverse types of services such as social media, news and mail.

One outcome is that multitasking is affecting the way users access sites and should be considered when measuring site engagement. We therefore define metrics that characterise multitasking during online sessions and show how they provide additional insights compared to standard site engagement metrics. Finally, we build patterns of multitasking that highlight the important multitasking characteristics of different sites. We discuss a number of insights on multitasking patterns, and show how these help to better understand how users engage with sites.

Section 5.2 discusses related work. Section 5.3 describes the data used in our study. Tree-streams are formally defined and studied in Section 5.4. The nature of multitasking is investigated in Section 5.5. The new metrics are defined and evaluated in Sections 5.6 and 5.7, respectively. In Section 5.8 and 5.9, we identify patterns of multitasking, and study their relationship. The chapter ends with a discussion.

5.2 Related Work

Online behaviour. How users arrive at or return to a page has been the focus of many studies *e.g.* [256]. For many years, users returned to a previously visited page via the back button [100]. Nowadays, the usage of tabs and tab switching has increased [105], and in fact overtaken the back button usage [70]. This way to navigate between pages is called parallel browsing or non-linear navigation, since pages can be (re)visited simultaneously. Tabs are particularly useful in online multitasking, as they allow users to pause one task to perform another [70; 258].

Different models have been proposed to describe this behaviour [25; 43]. In particular, referrer trees from server-side log data have been defined [177]. However, these cannot be used as they are to study engagement. Engagement metrics are calculated on a stream of interactions and there is no straightforward way to adapt them to the tree structure of more complex behaviours. In this chapter, we propose to “linearise” referrer trees into *tree-streams*, which we then use as our model of the user interaction data.

Multitasking. Online multitasking has been studied in the context of a web search session [170; 237]. For instance, it has been observed that multitasking happens in 81% of sessions [237]. Whereas a number of works provide evidence for multitasking during an online session [138], only Wang *et al.* [264] have studied this phenomenon in detail. Through an online survey, they showed that 92% of the participants had online sessions where they accessed several sites, to perform between 2 to 8 tasks.

Other works do not explicitly refer to online multitasking, but provide useful insights. For instance, users access different sites during a session [138] and a large proportion of pages are visited more than once (revisitation rate around 81% in 2001 [175] and 73% in 2005 [100]). In addition, the frequency at which a page is revisited differs depending on user habits and the type of website [138; 192], or, in other words, the web tasks a user accomplishes. With respect to the latter, three types of revisitation have been identified, short-term (backtrack, undo), medium-term (re-utilise, observe) and long-term revisits (rediscover), where it was shown that around 70% of revisits are short-term [100; 192].

All these provide a strong evidence that multitasking during online sessions exists and depends on the web tasks. However, since no metrics exist that explicitly account for multitasking, another focus of our work is the development of metrics that capture various aspects of online multitasking.

Table 5.1: Site categories and their subcategories, and percentages of sites in each (sub)category. For some subcategories a description is given as well.

Cat.	Subcat.	%Sites	Description
news 22.1%	news	5.79%	
	news (soc.)	5.13%	<i>society</i>
	news (sport)	2.63%	
	news (enter.)	2.24%	<i>music, movies, tv, etc.</i>
	news (finance)	1.97%	
	news (life)	1.58%	<i>health, housing, etc.</i>
	news (tech)	1.58%	<i>technology</i>
service 15.5%	service	7.63%	<i>translators, banks, etc.</i>
	mail	3.95%	
	maps	3.03%	
	organisation	0.92%	<i>bookmarks, calendar, etc.</i>
search 15.3%	search	12.63%	
	search (special)	1.58%	<i>search for lyrics, jobs, etc.</i>
	directory	1.05%	
sharing 9.6%	blogging	3.55%	
	knowledge	3.55%	<i>collaborative creation and collection of content</i>
	sharing	2.50%	<i>sharing of videos, files, etc.</i>
navi 9.3%	front page	6.58%	
	front page (p.)	1.84%	<i>personalised front pages</i>
	sitemap	0.92%	
leisure 8.7%	adult	2.76%	
	games	1.97%	
	social media	1.97%	
	dating	1.05%	
	entertainment	0.92%	<i>sites with music, tv, etc.</i>
support 8.7%	support	1.58%	<i>sites that provide products and support for them</i>
	download	7.11%	<i>downloading software</i>
shopping 7.9%	shopping	4.34%	
	auctions	2.11%	
	comparison	1.45%	<i>sites to compare prices of products</i>
settings 2.9%	login	1.71%	
	site settings	1.18%	<i>profile setting, site personalisation</i>

5.3 Dataset

We collected one month (July 2012) of anonymised interaction data from a sample of 2.5M users who gave their consent to provide browsing data through the toolbar of Yahoo. Users with very low or high activity (lower and upper 5% of the distribution) were excluded, which resulted in a dataset of 785M page views.

Site categories. We have shown in Chapter 4 that the type of site influences engagement. It is likely that online multitasking differs across sites of different categories, and that these differences impact the understanding and interpretation of common metrics that assess the online behaviour of users. We therefore selected the 760 most popular sites, measured by the number of users, in our dataset and annotated them using the categorisation schema of Section 3.3.

This resulted into a total of 11 distinct categories and 33 subcategories, which are listed in Table 5.1. The percentage of sites in these (sub)categories and a description of some of the subcategories are also shown. The categorised dataset contains 676 sites from 70 countries and regions, and accounts for 60% of the traffic in our original dataset. The sites cover a wide range of services (*e.g.* mail, news, shopping) sometimes catering for different subcategories of a given service (*e.g.* news about sport and finance).

5.4 Navigation Model

Previous work [100; 283] showed that users commonly use the back button to revisit sites and frequently maintain several tabs open and switch between them. We show that accounting for this behaviour provides additional insights about the effect of multitasking on metrics that measure online behaviour. To capture this type of navigation, we first define a new model called *tree-streams*.

Most studies investigating online behaviour model navigation as a linear click-stream; user interactions are ordered by timestamps and the accessed pages form a linear navigation path. This is illustrated in Figure 5.1 where (a) shows an example of log data and (b) shows the corresponding linear navigation path.

Click-streams based on standard server-side log data fail to capture some behaviours that we think are important when studying online multitasking or user engagement in general. Users may return to an open tab or window

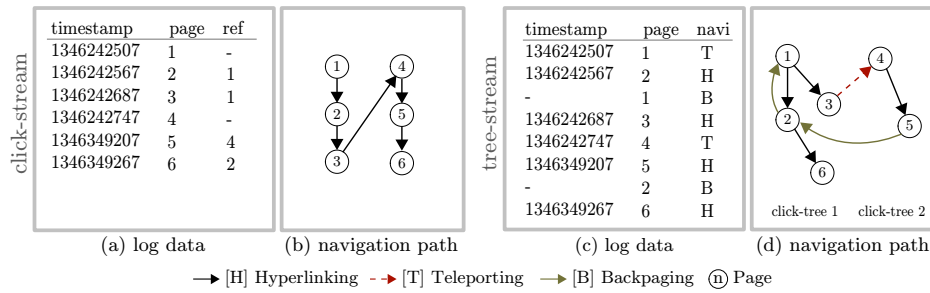


Figure 5.1: From click-streams to tree-streams.

or may use the back button to return to any previously visited and cached page. Since the revisit to a cached page does not require interaction with the web server, only the client-side log data records this type of navigation, and as such, is mostly non available. Therefore, we propose to model a part of client-side interactions using server-side log data, since these are more widely available.

Log data contain, beside the accessed page, the page the user is coming from (referral page): when a user is on a page and clicks on a hyperlink, the referral page is the page previously accessed. When no referral exists (as is the case with page 4), we deduce that the user jumped directly to the page using for instance a bookmark, a case of teleporting.

Now looking at Figure 5.1 (a), page 3 is accessed after page 2, but the referral is page 1. This implies that the user returned to page 1 before accessing page 3, probably using one or more tabs or the back button. The second visit to page 1 is not registered in the logs because browsers commonly cache the pages recently accessed to minimise bandwidth.

One way to consider the referrals when modelling user navigation are referrer trees, as proposed in [179; 257], where nodes represent pages and the links connect pages to their referrals. In Figure 5.1 (d) we have two referrer trees, one starting from page 1 and the other from page 4, as both pages lack a referral.

We “re-linearise” referrer trees into tree-streams as illustrated in Figure 5.1 (c) and (d) by re-introducing the missing referrals. This is necessary as engagement metrics for evaluating user behaviour are calculated on streams and not trees. Besides, processing trees typically requires higher time and space

complexity. Each pair of connected nodes is then labelled by one of the followings:

- **Teleporting [T]** Navigation without using a hyperlink (no referral is given) by entering a new URL, using a bookmark, *etc.*
- **Hyperlinking [H]** Navigation from one page to another using a hyperlink.
- **Backpaging [B]** Navigation to a cached page using tabs or back button. The page can be in the same referrer tree or part of another referrer tree.

As for click-streams, tree-streams are split into sessions, where a session ends if more than 30 minutes have elapsed between two successive page views. Finally, continuous page views of the same site are merged to form a site visit.

Tree-streams, as generated here from standard interaction data, do not contain all pages accessed via backpaging. A user may use the back button several times or return to several open tabs. Only the last page accessed in this manner and from which the user explicitly clicks on a link can be detected and included in the tree-stream. A more complete instrumentation would be required on the client side for all such pages to be detected.

Now we report some statistics. Compared with click-streams, tree-streams contain approximately 30% more page views. Moreover, 45% of the pages are accessed through hyperlinking, 31% through teleporting, and 24% through backpaging. In addition, 12% of the backpaged navigation land on a distinct tree, suggesting a task switch (or a different logical session according to the terminology in [179]).

Other studies reveal similar figures. For instance, Obendorf *et al.* report in [192] that 43.5% of navigation is via hyperlinking and 14.3% is via backpaging using the back button (their work did not consider backpaging with tabs). However, a study from Huang *et al.* [105] showed that 11.3% of the page views include tab switches. Combining back button and tab usage of the two studies suggests that 24% navigation happens through backpaging, comparable to our findings.

The proportion of teleporting reported in these studies is lower than what we observe in our dataset. However, these studies are based on a selective and comparatively limited set of 20 to 100 users. The log data used in our work includes millions of users with very different habits. Similarly to Kumar *et al.* [138] our dataset contains numerous sessions initiated by teleporting and only a few page views (see Section 5.5).

Backpaging and Dwell Time

A widely accepted metric is dwell time, which is the time users spend on a site during their visit. Calculating the exact dwell time is not obvious, since the time spent on the last page of a session is generally not known. Adding to this limitation, there is yet no consensus on how to identify when a session actually ends [176].

Multitasking makes it even more problematic to accurately calculate dwell time because of backpaging [121]. For instance, in Figure 5.1 the user backpaged from page 2 to page 1, from where he or she accessed page 3. No timestamp is associated with the user’s return to page 1, which makes it impossible to calculate the time spent on pages 1 and 2. To mitigate this problem we approximate the dwell time as described next.

Let i and j be two pages. Assume that a user backpages from i to j . The time spent on these two pages is known and can be written as:

$$t_{ij}^b = t_i^b + t_j^b$$

What we do not know is t_i^b and t_j^b , the time spent on each page, i and j , respectively. We propose to estimate these values by the time spent in each page when accessed via teleporting or hyperlinking, which will be generally known.

We denote these dwell time values t_i^* and t_j^* and devised three methods to estimate them:¹ we averaged the times of (1) all visits on pages i and j , (2) only those visits where pages i and j were accessed in the same order, and (3) only those sessions in which the site containing page i (or j) was visited at least twice in the same session. The second approach focuses on the same page visit pattern (i then j) whereas the third considers sessions on a site where multitasking occurs.

¹For simplicity, we defined the average dwell time over all pages of a site, not for each page separately.

We calculate for each approach the percentage (f) of cases for which t_i^* and t_j^* could not be computed because the pages were not visited through teleporting or hyperlinking. Since we analyze user online behaviour at site level, only backpaging between sites was taken into account. We restricted our analysis to sites viewed by at least 100 users. In total, we extracted around 17K page visit pairs.

We use linear regression to examine whether the estimations of t_i^b and t_j^b by t_i^* and t_j^* correlate with t_{ij}^b , using the linear equation:

$$t_{ij}^b = x_0 + x_1 \cdot t_i^* + x_2 \cdot t_j^*$$

A correlation of $r^2 = 0.43$ ($f = 0.00\%$) could be observed using all page views (approach 1), but the correlation increases ($r^2 = 0.50$, $f = 9.77\%$) when multitasking is considered (approach 3). The highest correlation ($r^2 = 0.52$) was obtained using the exact same pairs of page views (approach 2), but for $f = 29.89\%$ of cases t_i^* and t_j^* could not be defined. In all cases a low p-value was observed ($p - value \ll 0.01$).

The two coefficients are smaller than 1, but almost the same (*e.g.* for approach 2, $x_1 = x_2 = 0.69$) indicating that pages are visited in the same manner when using backpaging, but the time per page visit is smaller. This validates using t_i^* and t_j^* as estimate of t_i^b and t_j^b , respectively.

We use the third approach to estimate dwell time, because, although the correlation obtained is slightly lower than with the second approach, more dwell times could be approximated.

5.5 Characteristics and Effects

We present a number of characteristics of multitasking which we observed in our dataset. We also show how multitasking influences the way users visit a site.

5.5.1 Multitasking in Sessions

Our dataset contains 41 million sessions. The average number of sessions per user is 16.6 ($sd = 28.38$). Table 5.2 shows for sessions of increasing length, measured by page views, the multitasking and navigation statistics. The average values across all sessions appears in the last row. We define the degree of multitasking in an online session as the number of distinct sites visited during that session, denoted by $\#Sites$.

Table 5.2: Multitasking characteristics depending on the size of a session. Average and standard deviation are reported (*avg|sd*).

#PageViews \leq	#Sites	RecRate	Tele	Link	Back
2	1.10 0.30	0.00 0.00	—	—	—
4	1.42 0.78	0.01 0.00	40%	60%	0%
8	2.02 1.53	0.05 0.14	30%	52%	17%
16	3.01 2.79	0.10 0.17	29%	47%	24%
32	4.48 4.84	0.15 0.20	27%	45%	28%
64	6.38 7.94	0.19 0.24	25%	44%	31%
128	8.29 11.94	0.21 0.24	23%	44%	33%
256	9.62 15.90	0.22 0.24	22%	44%	34%
512	10.20 18.85	0.22 0.26	21%	44%	35%

In a session, a site may be revisited several times. Analogously to the page revisitation rate defined in [247], we define the site revisitation rate or *recurrence rate* as:

$$RecRate = \frac{\#Visits - \#Sites}{\#Visits}$$

where $\#Visits$ is the total number of site visits during the session. In addition, we provide the percentage of each navigation type in a session.

We can see that, on average, 10.20 ($sd = 18.85$) distinct sites are visited within a session, and for 22% ($sd = 0.26$) of the visits the site was accessed previously. The table shows that more sites are visited and revisited as the session length increases. Sessions with up to 16 page views consist on average of 3.01 distinct sites with a recurrence rate of 0.10. By contrast, sessions with up to 256 page views have on average 9.62 different visited sites with a recurrence rate of 0.22.

The way users revisit sites, whether via teleporting, hyperlinking or back-paging, also varies depending on the session length. Whereas teleporting and hyperlinking are the most important mechanisms to re-access a site during short sessions (we have 30% teleporting and 52% hyperlinking for sessions with ≤ 16 page views), backpaging becomes more predominant in longer sessions. Similar results were reported in [192]. Both, their and our study, show that tabs or the back button are often used to revisit a site, and that sites are often revisited during a session.

We look now at the session and site visit characteristics, presented in Table 5.3. We report for sessions of increasing length, measured by page views,

Table 5.3: Session and site visit characteristics depending on the size of a session. Average and standard deviation are reported (*avg|sd*).

#PageViews \leq	%Sessions	Session		Site visits			
		DwellTime [min]		#PageViews	DwellTime [min]		
2	27%	0.65	3.05	1.10	0.30	0.59	2.92
4	37%	2.14	6.15	1.28	0.65	1.51	4.72
8	50%	4.80	10.16	1.49	1.10	2.37	6.09
16	65%	8.67	15.26	1.69	1.69	2.88	6.99
32	79%	13.52	21.50	1.87	2.41	3.02	7.49
64	90%	18.72	28.60	2.03	3.34	2.93	7.70
128	97%	23.09	35.58	2.18	4.48	2.78	7.78
256	99%	25.59	40.68	2.29	5.78	2.66	7.84
512	100%	26.45	43.25	2.36	7.03	2.59	7.98

the percentage of sessions that have that length (*%Sessions*), and the dwell time time of the session (*DwellTime*). In addition, we measure the activity during a site visit with two metrics, page views (*#PageViews*) and dwell time (*DwellTime*).

We observe a relationship between the number of pages visited and the time spent on a page. For sessions with more than 32 page views, the number of page views per site visit increases with the session length, but the time spent on each visit to the site tends to decrease. In other words, the longer the session, and the more the users are multitasking, the quicker they navigate between pages, maybe skimming the content on the pages [79]. This results in more page views but a lower dwell time per site visit. In addition, backpaging increases with session length and is associated with shorter time spent on a page (see the coefficient of the linear regression model in Section 5.4). This suggests that visitors use backpaging to access previously visited pages or sites quicker, and spend less time on pages or sites they are returning to.

5.5.2 Visit and Inter-visit Activity

We have illustrated the importance of multitasking during online sessions and its relation with the navigation behaviour of users. In this section, we show how multitasking affects the browsing activity of users on a site, which we measure as the time spent on the site, aka dwell time. We do so for four selected categories of sites: news (finance), news (tech), social media, and mail. We extract for each category a random sample of 10,000 sessions.

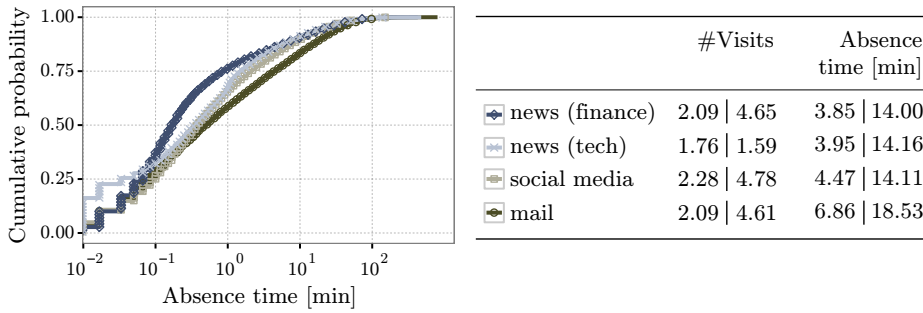


Figure 5.2: Site visit characteristics for four categories of sites: (Left) distribution of time between visits; and (Right) average and standard deviation of number of visits and time between visits ($avg|sd$).

Inter-visit activity. Figure 5.2 shows various statistics regarding the number of visits during a session and the time between visits. We refer to the latter as “absence time” following the study described in [72].

Sites with the highest number of visits within a session belong to the social media category ($avg = 2.28$, $sd = 4.78$), whereas news (tech) sites are the least revisited sites ($avg = 1.76$, $sd = 1.59$). These two categories have an average absence time of $4.47min$ ($sd = 14.11$) and $3.95min$ ($sd = 14.16$), respectively, although the distributions are similar.

The news (finance) sites have a skewer distribution, indicating a higher proportion of short absence time for sites in this category. Finally, mail sites have the highest absence time, $6.86min$ on average ($sd = 18.53$). However, when looking at the distributions of the absence time across all categories of sites, we see that the median is less than $1min$, and this for all categories. That is, many sites are revisited after a short break. We speculate that a short break corresponds to an interruption of the task being performed by the user (on the site), whereas a longer break indicates that the user is returning to the site to perform a new task.

Activity patterns. Next, we look at how multitasking is related to the way sites are revisited within a session. For each site, we select all sessions where the site was visited at least four times. Figure 5.3 (Top) shows the average dwell time at the i^{th} visit to a site, as a proportion of the total session length. The time spent on mail sites decreases at each revisit. The opposite is observed for social media sites. A possible explanation is that, for mail sites, there are less messages to read in subsequent visits, whereas

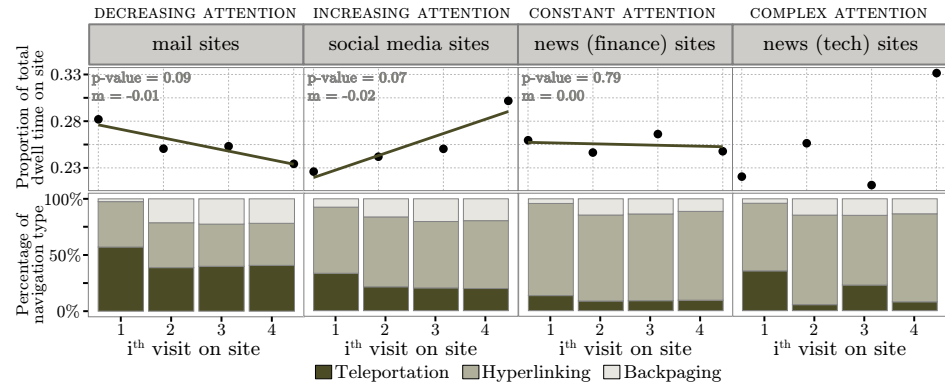


Figure 5.3: Activity patterns during online sessions: (Top) Visit activity described by the average dwell time of the i^{th} visit in a session. (Bottom) Usage of navigation types described by the percentage of each navigation type at the i^{th} visit in a session.

for social media sites, users might initiate new conversations with friends that are online.

News (finance) is an example of category for which neither a lower or higher dwell time is observed at each subsequent revisit. We hypothesise that each visit corresponds either to a new task or a user following some evolving piece of information such as checking the latest stock price figures.

Figure 5.3 (Bottom) shows how users access a site at the i^{th} visit as a percentage of the time they use teleportation, hyperlinking or backpaging. For all four categories of sites, the first visit is often through teleportation. Accessing a site in this manner indicates a high level of engagement, in particular in terms of loyalty, with the site, since users are likely to have bookmarked the site at some previous interaction with it. For instance, teleportation is more frequently used to access news (tech) sites than news (finance) sites.

After the first visit, backpaging is increasingly used to access a site. This is an indication that users leave the site by opening a new tab or window, and then return to the site later to continue whatever they were doing on the site. Finally, users still revisit a site mostly through hyperlinking, suggesting that links still have an important role in directing users to a site. For instance, news (finance) sites are mostly accessed through links; users are directed to sites of this category via a link.

The browsing activity for news (tech) sites (here measured by dwell time) is fluctuating. Either no pattern exist or the pattern is complex, and cannot easily be described. However, when looking at the first two visits or the last two visits, in both cases, a higher dwell time can be observed in each second visit. This may indicate that the visits belong to two different tasks, and each task is performed in two distinct visits to the site. Teleportation is more frequent at the 1st and 3th visits, which confirms this hypothesis.

We presented two main findings in this section. First, the time between two visits, or the absence time, can be used as an indication as to whether a user returns to a site to continue on a previously started task or to start a new task. The latter case is a sign of loyalty, as users return to the site to accomplish some new tasks.

Second, the activity pattern at the subsequent visits to a site provides additional information about how users engage with the site. We have identified four main patterns of *user attention to the site* (decreasing, increasing, constant and complex) and given examples of sites belonging to each in Figure 5.3. For instance, an increase in dwell time (increasing attention) can reflect “stickiness”: users are increasingly engaged with the site during the session. In the next section we define metrics that capture these and other multitasking characteristics.

5.6 Multitasking Metrics

In the previous section, we showed that on average 22% of the visits re-access sites previously visited within the same session, and that revisitation has an effect on user activity on the revisited sites. Therefore, only considering the user activity – in terms of dwell time and page views – within a visit provides a partial view of site engagement. In this section, we propose five multitasking metrics that cover different aspects of how users access a site during a session. All metrics are listed in Table 5.4.

Multitasking metrics. We employ two metrics that measure the extent of multitasking to a site. We know from Section 5.5.1 that users may return to a site several times during a session, and visit several other sites. To capture these, we define two measures: *SessVisits* is the average number of times the site was visited in a session and *SessSites* is the average number of sites accessed within a session. The latter metric measures the extent of multitasking with respect to how many other tasks (sites) are performed while visiting the site under consideration.

Table 5.4: Metrics used to analyse multitasking behaviour within sessions.

Metric	Description	Engagement	
		Low	High
Multitasking [MT]			
SessVisits	Avg. number of visits per session on site.	0	∞
SessSites	Avg. number of sites visited during the session where the site under consideration was visited.	0	∞
CumAct	Combining the dwell times of the visits on the site with accounting for the time between the visits.	0	∞
AttShift	Describes whether the dwell time on a site is decreasing, increasing or complex at each subsequent visit.	-1	1
AttRange	Describes the differences between the visits (dwell times) on a site in a session.	-	-
Activity [ACT]			
DwellTimeV	Avg. time per visit on site.	0	∞
DwellTimeS	Avg. time per session on site.	0	∞

To characterise the multitasking behaviour to a site we define three further metrics (*CumAct*, *AttShift*, *AttRange*) which we introduce in the following subsections. The metrics account for the browsing activity of users on a site which is commonly measured by the number of page views and dwell time. We focus on dwell time, but the metrics are easily adaptable to all measures that characterise the browsing activity on a site.

Activity metrics. We also consider two common metrics that measure the activity of users on a site. From Chapter 4 we use *DwellTimeV* which is the dwell time calculated at visit level. Additionally, we introduce *DwellTimeS* as the total time spent on a site during the full online session (see Table 5.4). Since *DwellTimeS* adds up the dwell times during an online session, it accounts in some way for the fact that users multitask when online. In the rest of this chapter, we compare our proposed metrics to these. We do not present the outcomes for the page view metric (*PageViewsV*), because the results are similar.

5.6.1 Cumulative Activity

Section 5.5.2 showed that looking at absence time provides some insights about how users engage with a site. We make the following assumption on how to interpret the time between site visits: if the next visit is shortly after the previous one, we consider that the two visits belong to the *same* task. If the time between the two visits is long, the user is returning to the

site to perform a different task.² In case of a search engine for example, a short absence time could refer to the same search task, whereas a long one indicates a different search. The latter case is related to the loyalty of users to a site, whereas the former is more related to the actual activity of the user on the site, an activity that was briefly interrupted. Drawing from the web search context again [111], a short absence time – often taken to be less than 30 minutes in the literature – is indicative of a query re-formulation or a re-orientation of the original task, whereas a long absence time indicates that the two sessions are unrelated.

We use the following metric to express this:

$$CumAct_k = \log_{10}(v_1 + \sum_{i=2}^n v_i \cdot iv_i^k)$$

Here v_i corresponds to the dwell time during the i^{th} visit, iv_i is the time between the $(i-1)^{th}$ and i^{th} visit, and n is the total number of visits. With our definition of $CumAct_k$, a long absence time between visits to a site is an indication of a high loyalty to that site: the user is returning to the site to perform some new tasks. The value of $CumAct_k$, referred to as the *cumulative activity* metric, increases with the time spent between visits to the site.

The exponent k is used to scale the iv_i parameter; a high value increases the importance of between-visit activity iv_i . When $k = 0$, the focus is solely on the visits to the site, and $CumAct_0$ corresponds to the total time on the site during the whole session (as in *DwellTimeS*).

Selection of k . The exponent k can take several values. To verify that the proposed cumulative activity metric adds new insight, we must ensure that it does not capture the same information as the common metrics *DwellTimeV* and *DwellTimeS*. We could compare the scores the different metrics attribute to different sites, but it is unclear how these scores should be normalised to be comparable. We therefore compare instead the rankings of sites using *DwellTimeV*, *DwellTimeS* and $CumAct_k$ for various values of k with the Kendall tau coefficient (τ). In addition, also using τ , we compare the rankings of sites using two successive values of k , that is $CumAct_k$ and $CumAct_{k-0.5}$. The objective is to determine the extent to which accounting for absence time affects the ranking of sites; a high τ value means that the sites are ranked similarly even though the importance associated with the absence time has increased by 0.5. All comparisons are shown in Figure 5.4.

²We postpone for now the problem of defining short and long absence time.

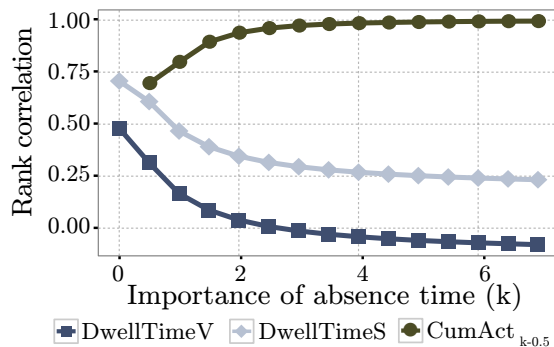


Figure 5.4: Cumulative activity for different importance values of the absence times (k).

While the value of k is low, corresponding to a low importance given to absence time, the cumulative activity metric correlates highly with *DwellTimeV* and *DwellTimeS*. The measurement is dominated by site visits. Increasing k to 3 causes τ to decrease (for instance we see that $\tau(\text{CumAct}_3, \text{DwellTimeS}) = 0.30$), whereas the correlation between the rankings defined by successive values of k and $k - 0.5$ increases (we have $\tau(\text{CumAct}_3, \text{CumAct}_{2.5}) = 0.97$). This shows that accounting for absence time captures the effect of multitasking when measuring user browsing activity. However, values of k higher than 3 lead to minor differences in the τ values, indicating that absence time does not bring new perspectives on multitasking. We therefore fix k to 3 and simplify the metric name to *CumAct* for the rest of our study.

5.6.2 Activity Patterns

We know from Section 5.5.2 that the browsing activity during a site visits varies and that these variations depend partly on the type of sites (*e.g.* news vs. mail). Motivated by this, we define two measures that describe how the dwell time is changing from visit to visit.

We interpret an increase in dwell time as a user getting increasingly more engaged with the site at each revisit, whereas we hypothesise that a decrease in dwell time corresponds to a user shifting his or her attention away from the site, arguably because the focus moves to some other task on another site. Finally, a constant dwell time is interpreted as the user repeatedly visiting the site to perform the same type of task. We do not attempt to automatically identify which patterns apply to which types of sites (we leave

this for future work); our focus is the provision of measures that account for such patterns.

We modify the measure of “inversion number”, a common measure of sort-
edness, to express the above:

$$inv_n = \sum_{i,j}^n \{v_i + v_j \mid i < j \text{ and } v_i \leq v_j\}$$

Here v_i and v_j with $i < j$ correspond to the browsing activity during the i^{th} and j^{th} visit, respectively. Whereas the original inversion number determines the number of (v_i, v_j) pairs that do not exhibit a natural order ($v_i > v_j$), our measure counts how often an increase or no change in browsing activity is observed ($v_i \leq v_j$). Moreover, inv_n considers the extent to which the browsing activity changes when comparing the i^{th} to the j^{th} visit of a site, where, for instance, an increase of dwell time from $v_i = 10sec$ to $v_j = 12$ secs is considered less important than one from $v_i = 10$ secs but to $v_j = 2min$. We use n to refer to the number of visits to a site within a session. As shown in Section 5.5.2, how users engage with a site during an online session depends on how often the site is revisited.

We next normalise inv_n between -1 and 1 to define a measure that models the shift of attention in the browsing activity during a session:

$$AttShift_n = 2 \frac{inv_n - minInv_n}{|maxInv_n| - |minInv_n|} - 1$$

Here $minInv_n$ and $maxInv_n$ are the inversion numbers after re-ordering the site visits according to, respectively, decreasing and increasing values of the dwell time on the site. When $AttShift_n$ equals to 1 , the attention is shifting towards the site (*increasing attention*). When $AttShift_n$ equals to -1 , we have the opposite, the attention is shifting away from the site (*decreasing attention*). Finally a value of $AttShift_n$ close to 0 means no identifiable patterns; we refer to this as *complex attention*.

Finally, the browsing activity at each visit may remains constant, *e.g.* same dwell time at each visit. $AttShift_n$ cannot capture this, so we define an additional measure to express this:

$$AttRange_n = \frac{\sigma(V_n)}{\mu(V_n)}$$

where $\sigma(V_n)$ is the variance and $\mu(V_n)$ is the average dwell time of all visits V on a site during a session. $AttRange_n$ is a normalised variance and a value

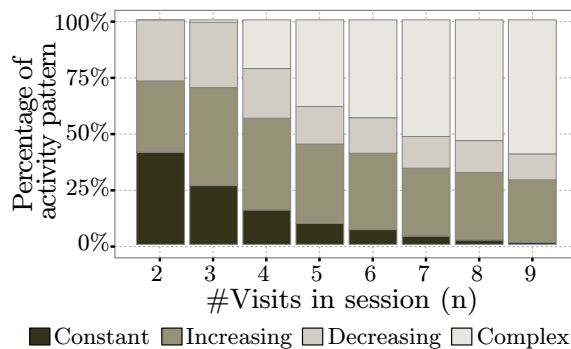


Figure 5.5: Activity patterns for different numbers of visits in a session (n).

near 0 indicates that the browsing activity exhibits only small fluctuations; we refer to this as *constant attention*.

Selection of n . We aim to understand how the activity pattern changes depending on the number of visits in a session. First, we use $AttRange_n < 0.1$ to express that the attention remains constant. For all values above 0.1, we use $AttShift_n$ to identify the type of activity pattern, namely, increasing, decreasing or complex attention. We say that $AttShift_n < -0.25$ indicates a decrease of attention, $AttShift_n > 0.25$ means that the attention is shifting towards the site at each subsequent visit, and any value between -0.25 and 0.25 indicates a complex attention. The values 0.1 and 0.25 are chosen arbitrarily, and are sufficient to analyse the effect of n on the activity patterns.³ Figure 5.5 reports the percentage of sites that belong to each type of activity patterns for different number values of n .

The more often sites are visited in a session (the higher the value of n), the less their activity pattern is either constant or decreasing. In addition, the number of sites that follow an increasing attention pattern does not decrease as much. The percentage of sites with a complex activity pattern however increases. We hypothesise that this is because the site is visited to perform separate tasks, but when a task is carried out in several visits, these visits may follow some specific (not complex) activity patterns (as illustrated with news (tech) sites in Figure 5.3). In the rest of this chapter, we fix n to 4, as the sites are distributed equally over the four types of activity patterns. Additionally, we simplify the metric names to $AttShift$ and $AttRange$.

³For instance, a higher value than 0.1 would lead to more patterns classified as constant. We leave for future work the detailed study of activity patterns.

5.7 Evaluation

The newly proposed multitasking metrics, *SessVisits*, *SessSites*, *CumAct*, *AttShift*, and *AttRange* provide new insights about the online behaviour of users. We contrast these with the site engagement metrics *DwellTimeV* and *DwellTimeS*.

To compare metrics, we rank sites according to each metric and we then evaluate the similarity between these rankings. Admittedly, if two metrics produce the same ranking, they are equivalent and hence redundant. The Spearman's rho coefficient (ρ) is used to compare the rankings. We only consider correlations that are statistically significant (p-value < 0.05). The results are reported in Table 5.5.

There are no or only weak correlations between the multitasking metrics (all correlations are below 0.45). We even cannot report the correlation coefficient between some metrics, since the p-value is above 0.05. We observe a weak correlation between *SessVisits*, and *SessSites* and *CumAct*, indicating that the more often users return to a site during a session, the more other sites are accessed ($\rho(\textit{SessVisits}, \textit{SessSites}) = 0.42$) and the more time users spend on the other sites ($\rho(\textit{SessVisits}, \textit{CumAct}) = 0.41$).

The weak negative correlation ($\rho = -0.38$) between *CumAct* and *AttRange* suggests that if the time between visits is high (*i.e.* each visit refers to a new task), a similar dwell time at each subsequent revisit can be observed. This is in accordance to the results of Section 5.5.2, where we could see that visits belong to different tasks, if the activity remains constant.

However, the correlations are not high enough to suggest that the metrics are redundant. We therefore conclude that all multitasking metrics convey different information about the online behaviour of users.

We can also see that the multitasking metrics rank sites differently compared to the two activity metrics (all correlations are below 0.4). This implies that accounting for multitasking, with the new metrics *SessVisits*, *SessSites*, *CumAct*, *AttShift*, and *AttRange*, leads to different conclusions with regard to engagement during a visit (*DwellTimeV*), or across the entire session (*DwellTimeS*).

We only observe a weak negative correlation between *SessVisits* and *DwellTimeV* ($\rho = -0.40$) suggesting that a high number of visits to a site can lead to a lower dwell time per visit. We speculate that the visits belong

Table 5.5: Spearman’s rho between multitasking and activity metrics. Correlations with a p-value ≥ 0.05 are not reported (-).

	[MT] SessVisits	[MT] SessSites	[MT] CumAct	[MT] AttShift	[MT] AttRange	[ACT] DwellTimeS
SessSites [MT]	0.42					
CumAct [MT]	0.41	-				
AttShift [MT]	0.09	-	-			
AttRange [MT]	-	-	-0.38	0.27		
DwellTimeS [ACT]	0.20	0.24	0.12	0.32	0.08	
DwellTimeV [ACT]	-0.40	-	-	0.14	-	0.50

to the same task, and hence the dwell time required to perform the task is distributed over several sub-visits.

The metrics *DwellTimeV* and *DwellTimeS* have the strongest positive correlation ($\tau(DwellTimeV, DwellTimeS) = 0.50$). However, considering that both metrics characterise the activity on a site (per visit and session, respectively), the observed correlation is rather low. This shows that the multitasking effect is significant, because not accounting for multitasking (*DwellTimeV*) leads to a different ranking of the sites as when accounting for it (*DwellTimeS*).

We also compared the ranking of sites per category. For instance, search sites and social media sites provide different services, and it is important to know how the same type of service is engaged with by users across the sites offering that service. We can report that the average rank correlation per category is between $\rho = 0.08$ and $\rho = 0.30$ which shows that the metrics also differ when comparing the ranking of sites of the same category.

5.8 Multitasking Patterns

We examine the different multitasking patterns that sites exhibit. Our objective is to analyse how users engage with a site, in terms of their browsing activity, and how multitasking influences it. To identify the multitasking patterns, we cluster sites characterised by their *DwellTimeV*, *DwellTimeS*, *CumAct*, *SessVisits*, and *SessSites* values. The k-means algorithm is used to perform the clustering. Since our metric values do not follow a normal distribution, thus to avoid the extensive influence of heavy outliers, we do not use the value of each metric, but the corresponding site rank. The number of clusters is determined by a minimal cluster size such that each cluster contains at least 20% of the sites.

Five clusters (patterns) were identified, shown in Figure 5.6, each representing a set of sites that exhibit a similar multitasking and visit activity. The first row of the figure contains the cluster centroids normalised by the z-score. Each bar corresponds to one metric. The vertical axis shows how many standard deviations a rank value is above or below the mean rank. This means that bars above zero indicate higher ranks for the respective measure whereas bars below zero indicate lower ranks. The second row contains the number of sites within each cluster, and the third row presents the predominant site categories per cluster. The probability difference *PD* corresponds to the likelihood that a category occurs in a given cluster with respect to its likelihood that it occurs at all (see Section 3.3 for a detailed definition of *PD*). The last row of Figure 5.6 reports the percentage of the activity patterns per cluster, measured by the metrics *AttShift* and *AttRange*. Each cluster is given a name reflecting its main characteristic.

We can see that the first two patterns are characterised by a single-task-oriented browsing activity. Users do not access several sites within an online session (low *SessSites* and *SessVisits*), implying that they perform only one task, on the site under consideration. The other three patterns describe multitask-oriented browsing activity where users access several sites (high *SessSites*) and re-access the focal site several times (high *SessVisits*) within an online session. We discuss now the identified multitasking patterns in detail and relate them to different types of tasks performed on sites.

Quick tasks. Sites following this pattern are characterised by a short dwell time per visit (*DwellTimeV*) and over the whole session (*DwellTimeS*); users do not spend a lot of time on these sites. Moreover, users visit these sites to perform a single task – without interruption – (low *SessVisits*), and

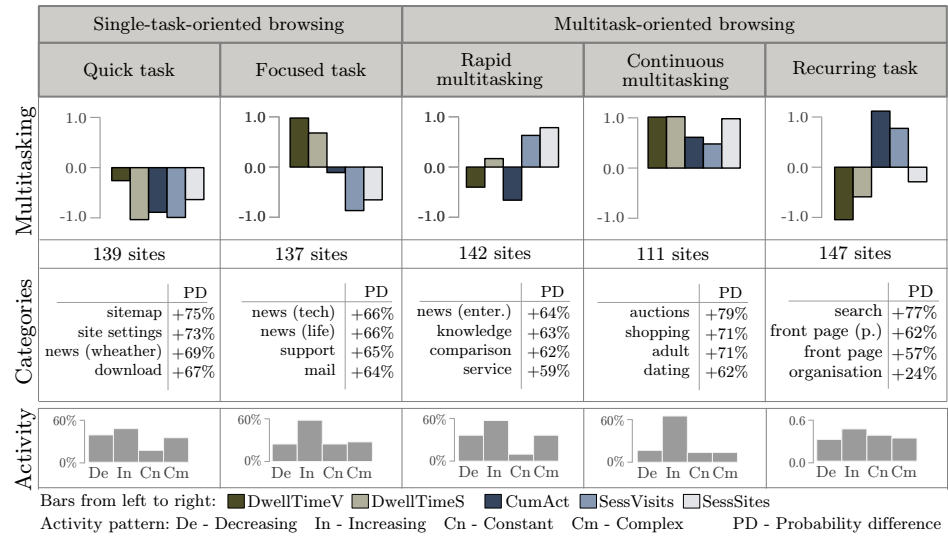


Figure 5.6: Multitasking patterns during online sessions: (1st row) Site clusters and browsing characteristics. (2nd row) Number of sites per cluster. (3rd row) Predominant site categories per cluster. (4th row) Percentage of activity patterns per cluster.

they do not return after a longer time of absence to perform new tasks (low *CumAct*). News (weather) and download sites belong to this cluster. These type of sites are visited to perform a precise and quick task, *e.g.* checking the weather and hence the results intuitively make sense.

Focused tasks. Sites belonging to this pattern (*e.g.* news (tech and life), mail) have a high activity per visit (*DwellTimeV*) and during the whole session (*DwellTimeS*). In addition, the low values of *SessVisits* and *SessSites* indicate that the focus of the users is solely on the site under consideration. This means that users visit the site to perform their task (*e.g.* reading news or mails) during a single visit, and they also do not get involved in any other tasks during that online session.

Rapid multitasking. Sites belonging to this pattern (*e.g.* news (entertainment), knowledge, comparison) are sites for which the multitasking effect is significant. The dwell time per visit (*DwellTimeV*) is lower than on average, but the dwell time per session (*DwellTimeS*) is higher. This means that measuring engagement at visit level only can lead to incorrect conclusions, because users return several times to the site during a session (high *SessVis-*

its). The low value of *CumAct* indicates that users quickly return to the site under consideration, probably to continue with the same task. This implies that the task on the site is actually split into several sub-visits, whereas the time between two visits is very short. This is further accentuated by the low percentage of constant activity patterns (10%), which suggests that the visits are connected with each other, because the attention does not remain the same. Instead, users likely shift their attention “towards” or “away” from the site at each subsequent visit (58% of the sites have an increasing, and 37% of the sites have a decreasing activity pattern, respectively). For 37% of the sites the activity pattern is complex: only parts of the visits belong to a same task.

Continuous multitasking. This pattern refers to sites that are continuously visited by users, even after a longer absence time (high *SessVisits* and *CumAct*). The users spend much time on the sites during each visit (high *DwellTimeV*). Many auctions and shopping sites belong to this cluster. We hypothesise that, for instance, online shoppers take some time before they decide to purchase an item. In doing so, the shopping task is actually split into several sub-visits whereas the time between two visits can be long. In the meantime (*i.e.* between two visits) users might even check offers from other sites or read reviews about the product [208]. In the case of auction sites, users actually need to return regularly to follow the auction and maybe to bid again on an item they want to purchase. We also observe that for 66% of the sites the dwell time increases at each subsequent visit (increasing attention pattern), which suggests that the focus of the user is moving towards the site (*e.g.* users want to make some purchases on a shopping site, hence becoming more focused in this task).

Recurring tasks. This pattern primarily contains search sites and front pages that are characterised by the highest average number of visits per session (*SessVisits*) and the highest cumulative activity (*CumAct*) compared to the other patterns. This shows that users regularly return to the sites and that they perform several (probably recurrent) tasks within a session (*e.g.* they have several search tasks). Looking at the activity patterns of the cluster we can see that there is no dominant pattern, as all four patterns occur to the same extent in this cluster. We speculate that users either return to the site after a short absence time to continue with the same task (*e.g.* they continue with the same search task), or they return later to perform a new task (*e.g.* they start a new search task). In the former case, the dwell time can decrease (users navigate quickly to the next search result)

or increase (users reformulate the query) at each subsequent visit, whereas in the latter case, the dwell time remains the same (users dedicate the same time for each search task) or fluctuates (each search task has several visits).

In this section, we showed that our proposed metrics reflect how users engage with a site, in terms of their browsing activity during the entire session, in the presence of multitasking. We identified different patterns of multitasking and observed that the patterns differ significantly between sites.


Some sites are visited once to perform a single task and depending on the task users spend much or little time on the site. Other sites are characterised by steady return of users even after long absence times. In this case each visit usually corresponds to a new task, but the visits can also be connected with each other. For the latter case, user attention shows a particular trend: it is shifting towards the site.

Finally, there are sites for which the multitasking effect is significant. The task on the site is repeatedly briefly interrupted by visits to other sites resulting in a low activity per visit but a high activity over the whole session. Also for this multitasking pattern, it is likely that the focus of the user is moving towards the site.

5.9 Relationship between Patterns

The conducted clustering groups sites that exhibit a similar browsing and multitasking activity. We study now whether sites of the same category (*e.g.* all news sites) belong to the same cluster, that is, whether they follow the same multitasking pattern. Figure 5.7 shows the percentage of sites of a category belonging to a cluster. Each column corresponds to a site category and the colours in the column represent how the sites of that category are distributed over the multitasking patterns (clusters). Additionally, we identify for each site category its “home” cluster, which is the multitasking pattern that most of the sites of that category follow. Site categories (columns) that have the same “home” cluster are grouped together.

We can see that there are categories where almost all sites of the category belong to the same cluster. For instance, 79% of the auction sites follow the “continuous multitasking”, and 93% of the search sites follow the “recurring task” pattern. We can report that on average 52% (sd=14%) of the sites of a category belong to their “home” cluster, and the remaining sites are distributed over the other clusters.



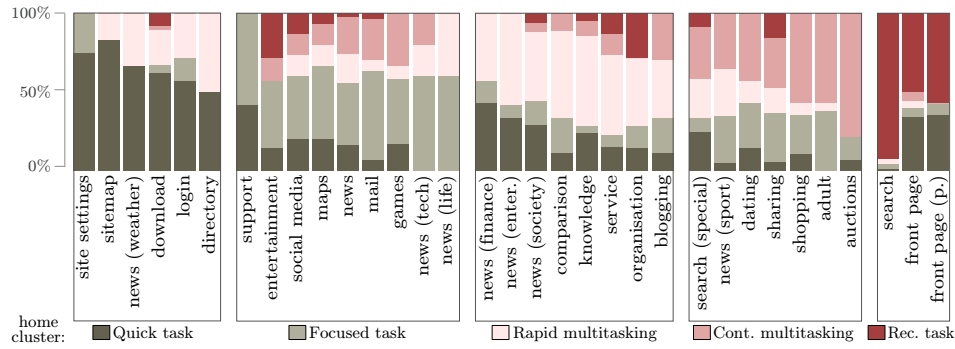


Figure 5.7: Distribution of sites of a category over the clusters, and the “home” cluster (*i.e.* multitasking pattern) for each site category.

Then, there are categories where the browsing activity of the corresponding sites differs. For instance, news sites mainly belong to the “Focused tasks” cluster, but there are also many news sites that follow the continuous or rapid multitasking pattern. Indeed, the sites of a category belong to on average 3.76 ($sd=1.12$) different clusters. The overlaps between clusters and site categories might point to hidden dependencies between the clusters. For example, we can see that if most of the sites of a category (*e.g.* shopping) follow the “Continuous multitasking” pattern, other sites of the same category tend to follow the “Focus task” pattern.

This dependency is further accentuated when looking at the browsing characteristics of the patterns (see Figure 5.6). For both patterns the dwell time per visit is high and increasing at each subsequent visit (increasing activity). Only the multitasking characteristics of the two patterns differ. The latter pattern describes single-task-oriented browsing (*e.g.* users search for items they want to purchase), whereas the former exhibits multitask-oriented browsing (*e.g.* users also return to make a purchase). In addition, the multitask-oriented pattern have a higher dwell time across the whole session. Similar observations can be made when comparing the “quick task” and “rapid multitasking” patterns.

In summary, sites that exhibit multitask-oriented browsing can be more “sticky” than sites that are characterised by single-task-oriented browsing. Although all sites have a high dwell time per visit, users return several times to the former sites – even after a longer absence time – which results in a higher overall dwell time during a session.

5.10 Discussion

This chapter studied online multitasking and its effect on measuring site engagement. We focused on one aspect of engagement, “stickiness”, which is concerned with users spending time on a site. Stickiness is mostly measured with metrics assessing users’ depth of interaction with a site. We focused on two such metrics, page views and dwell time. Our study is based on the online interactions of a large sample of users and their online browsing activity on 676 sites.

Most studies investigating online behaviour model user navigation with linear click-streams. Users may return to a site via an open tab or window or use the back button to return to any previously accessed page, from which they access the next page. Both cases generate additional page views that are not part of standard linear click-streams. Referrer-trees overcome this but cannot be used in a straightforward way to calculate engagement metrics. We “re-linearised” referrer-trees into tree-streams, which are like click-streams, but offer a richer representation of the interaction data. We are able to study how users visit websites, whether through teleporting, backpaging or hyperlinking, bringing additional insights about online multitasking.

We have shown that online multitasking exists, as many sites are visited and revisited during a session. We also demonstrated that multitasking influences the way users access sites and that this depends on the site under consideration. Metrics that describe the browsing activity - such as dwell time and page views during a single visit or the whole session - do not account for multitasking, that is how the visits differ from each other and what users are doing while not on the site. Therefore, we defined five new metrics, each aiming at capturing specific aspects of the browsing activity with a site during a session when multitasking is involved.

We employed two metrics that measure the extent of multitasking to a site. The metric *SessVisits* measures how often users return to a site, whereas the metric *SessSites* characterises how many other tasks (sites) are performed while visiting the site under consideration. The cumulative activity metric *CumAct* accounts for the activity between site visits. The longer the time between two visits, the higher the likelihood that the user is returning to that site to perform a new task. This somewhat reflects the loyalty of the user to the site, and the cumulative activity metric increases the importance of visits that preceded a longer break of activity on that site.

The activity pattern metrics, attention range *AttRange* and attention shift *AttShift*, provide information about what is happening on a site at each revisit. These metrics allow us to identify the sites for which the attention of the user is “shifting” towards or “away” from the site. In the former, the user becomes more focused in the task being performed by visiting that site, whereas in the latter, the user is slowly doing something else, on some other sites.

All metrics were compared, in terms of how they correlate, and we could clearly see that our proposed metrics indeed provide different insights about how users engage with a site at the visit-level and session-level, acknowledging that users are going to other sites during a session.

We know from Chapter 4 that sites of the same category can differ in their engagement characteristics. Although the analysis of this chapter is based on a much more fine-grained categorisation schema (33 instead of 6 categories), the same observations could be made. As already discussed in Chapter 4, these differences are probably a consequence of structural and content-related differences between the sites. However, some design decisions might lead to a more engaging user experience than others.

Finally, the results showed that leaving a site does not necessarily entail less engagement, as users often switch between sites and hence leaving them but returning later on. If the sites belong to the same provider, it is even desirable that users switch between them, as it implies that the users engage with more than one site of that provider. How to measure engagement within a network of sites, and accounting for the traffic (interactions) between sites, is the focus of our next chapter.

Concluding remarks. In the remaining chapters, unless otherwise stated, we model user online behaviour as tree-streams, and we then measure the browsing activity on a site at session- instead of visit-level. Both approaches enable us to obtain a more accurate picture regarding the browsing activity of users on a site, and hence their engagement with that site.

Limitations. The work of this chapter has two main limitations. First, our definition of a task is simplistic and should be extended. For example, to book a holiday, a user may visit several sites within the same session and one site many times across several sessions. Moreover, we need to further study when a user is returning to a site to continue with the same task (continue reading the news articles) or to start a totally different task

(a new unrelated information need), as these two cases should be treated differently. Accounting for this will likely lead to more advanced models of user attention with a site, an important research direction to follow.

Second, the log data used in this study fail to capture some user interactions that are important. Although tree-streams give a better overall picture about how users navigate through the Web, only log data that record all types of interactions with the web browser (*e.g.* tab and back button usage) are able to capture the whole navigation path of users. This also implies that the real extent of online multitasking is even higher than measured in this chapter. However, these data are mostly non available, especially at a larger scale, and therefore could not be used in this study.



Inter-site Engagement

This chapter proposes an approach for measuring engagement with a network of websites by accounting for the traffic between the sites. We refer to this type of engagement as *inter-site engagement*.

6.1 Introduction

Chapter 5 provided evidence that users multitask within their online session. Online multitasking implies that users regularly visit several sites during an online session, and that many of these sites are visited more than once. Taking this into account, it might be interesting to study how users engage across sites when performing a single task (*e.g.* shopping, news reading) and how this affects the engagement with each site.

In addition, many large online providers (*e.g.* Amazon, Google, Yahoo) offer a variety of sites, ranging from shopping to news. 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 with 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 site engagement.

Moreover, site engagement metrics are not applicable when assessing engagement with more than one site, as they do not account for the user traffic between sites. We therefore propose a methodology for studying *inter-site engagement*, that is, 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 in conjunction with site engagement metrics to measure the engagement with respect to a 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 [190], for instance, density at network-level and page rank at node-level. In this chapter, we demonstrate the value of our approach on a provider network of 155 sites offered by Yahoo.

We start by covering previous work in Section 6.2. Section 6.3 describes the data and network instances used in our study. The characteristics of the networks are described in Section 6.4. The inter-site metrics used in our work are introduced and evaluated in Sections 6.5 and 6.6, respectively. We then apply the metrics in Sections 6.7 and 6.8, and define patterns of inter-site engagement. The chapter ends with a discussion.

6.2 Related Work

Online behaviour. User online behaviour on the Web has been studied in a number of contexts, for example looking at the general browsing characteristics of users [49; 138; 179], how users visit websites [32], the return rate to a website [72], or how users discover and explore new sites [18]. From these and other studies, several user navigation models were developed [44; 179; 232], 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 [179; 232], and loyalty to a site [18; 49].

Several studies focused on the online behaviour across sites (*e.g.* [36; 104]), and how to support such online behaviour using cross-site personalisation [131; 265]. Other research investigated how the fact that users engage with several sites on the Web can be used to develop cross-domain recommendations, that is, using the experience of a user on one site to make recommendations on another site [76; 101]. There is research [114; 115; 204] that 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 [114], and that the more active shoppers tend to visit more shopping sites [115]. Other research studied the reading behaviour across news sites [36; 220]. A recent study found that 57% of users routinely obtain their news from between two to five news sites [36]. Moreover, social media and search sites are playing an important role in the consumption of news [37].

Network analysis. In this chapter, 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 [190] together with engagement metrics to study *inter-site engagement*.

Many types of complex networks have been studied [190]; those closer to our work are user traffic networks [44; 178; 275] investigating the traffic between web elements (pages, hosts, and sites), and hyperlink networks [17] 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 [44; 178]. Some studies also considered engagement metrics [44].

There are many ranking algorithms that incorporate the online behaviour within a network or its hyperlink structure to rank pages [168; 202], images [251], news articles [252], *etc.* The most famous ranking algorithm is PageRank [202], which ranks pages using the structure of the hyperlink network. However, recent research [168] 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 network according to how much users engage with the network after visiting that site. This can be compared to the problem of influence maximisation [41; 42; 125], *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 with a network of sites. The measure itself is an adaption of the downstream metric of [280] to our network model.

Table 6.1: Network instances based on five country in one continent. For each network, we provide the number of network instances, and the average and standard deviation of the number of clicks per instance.

Type	Network	Number of instances	Clicks per network
Month	YH_{Feb}	1	16M
Daily	$YH_{01/08}, \dots, YH_{31/07}$	356	577K 230K
Country	YH_{c1}, \dots, YH_{c5}	5	3.2M 3.6M

6.3 Dataset and Networks

Our study is based on anonymised interaction data 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. We extracted a total of 53M sessions. The browsing activity of a user on Yahoo during a session was used to create several provider (in our case Yahoo) networks as described next.

Provider networks. Our aim is to provide insights about user engagement with respect to Yahoo’s network of sites. We created a provider network using a total of 155 sites from five countries from the same continent. We selected the 31 most popular sites that have a counterpart in all of the countries, and that encompasses diverse services such as news, mail, and search.¹ The provider network is a weighted directed network $G = (N, E)$, 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 in our data. This allows us to create different network instances.

Network instances. The network instances are listed in Table 6.1. We defined various network instances of the provider network to evaluate the metrics described in Section 6.5. The metrics characterise the inter-site engagement between nodes or within the whole network. The metrics are evaluated in Section 6.6. We extract browsing data from each day between August 2013 and July 2014. The browsing data are used to create 365 networks ($YH_{01/08}, \dots, YH_{31/07}$), each representing the traffic of the network

¹We consider all subdomains in Yahoo (*e.g.* mail.yahoo.com), and other domains that belong to Yahoo (*e.g.* flickr.com).

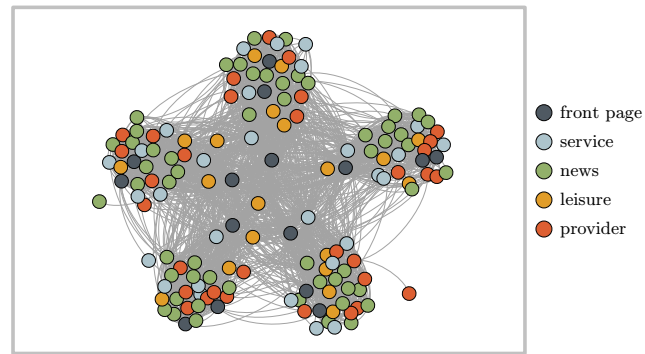


Figure 6.1: The provider network based on browsing data of February 2014. The edges are not weighted for confidentiality reasons. The Force-Atlas algorithm is used for the visualisation.

on a certain day (here 08 refers to August, 07 to July, *etc.*). We also define a network YH_{Feb} based on the browsing data of February 2014 to study the differences of inter-site engagement between sites. Finally, five country-based networks are created (YH_{c1}, \dots, YH_{c5}), where each contains the sites and traffic of one country of the provider network. Looking at specific countries enables us to compare the inter-site engagement, on a per country basis.

Site categories. We use the categorisation schema of Section 3.3 to annotate the sites of our provider network. This results in the following site categories:

- 35% news sites (*e.g.* news, finance) [News]
- 19% service sites (*e.g.* mail, calendar) [Service]
- 13% leisure and social media sites (*e.g.* tumblr, games) [Leisure]
- 23% provider sites (*e.g.* account settings, help) [Provider]
- 10% front pages and site maps (*e.g.* My Yahoo, Yahoo Everything) [Front page]

An example of the provider network is displayed in Figure 6.1. The nodes represent the sites colored by their category. 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.

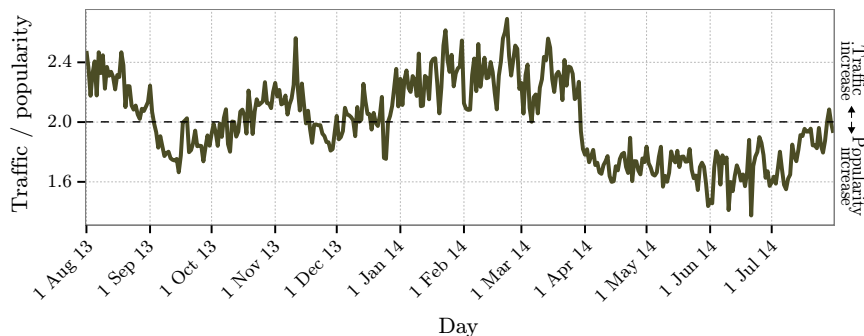


Figure 6.2: Fluctuations of network popularity and traffic over time.

6.4 Characteristics and Effects

We present a number of characteristics of our network. We also show how the traffic in the network affects the popularity of sites (nodes).

6.4.1 Traffic and Popularity

We focus on the dependencies between the traffic and the popularity of networks, and the popularity of sites and the traffic between them.

Networks. We use the daily networks introduced in Section 6.3, namely $YH_{01/08}, \dots, YH_{31/07}$. We calculate for each network (day) the *popularity* of the network measured by the number of users that visited the network on that day, and the *traffic* in the network measured by the total number of clicks between nodes (*i.e.* the sum of the edge weights).

Figure 6.2 presents the daily traffic divided by the daily popularity. The average is 2, *i.e.* there is twice as much traffic in the network than users. This implies that users regularly navigate between the sites in the network. However, we also observe fluctuations over time. For instance, between January 2014 and April 2014, the value is above average indicating that there was an increase in traffic, but the number of users remains the same. On the other hand, between April 2014 and July 2014, the value is below average indicating that the traffic decreases.

Overall, this shows that the traffic and popularity of a network do not always change in the same manner. Hence, accounting for the traffic in the network reveals new insights about inter-site engagement, as some networks might

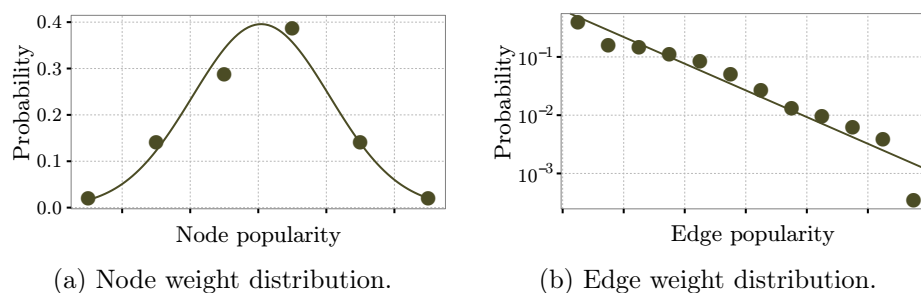


Figure 6.3: Distributions of node weights (popularity) and edge weights (traffic). The x-axis values (log-scale) are removed for confidentiality reasons.

have a high popularity (many users), but low traffic (users do not navigate between sites), and vice versa.

Nodes and edges. We investigate now the characteristics of the nodes and edges of the network. We use the network constructed for February 2014 (YH_{Feb}). We calculate the popularity of each node (site) as the number of users that visited that site. The edge popularity is described by the edge weight, hence, a higher weight corresponds to a higher popularity.

Figures 6.3a and 6.3b present the distribution of node and edge popularity, respectively. We observe that node popularity follows a lognormal distribution, indicating the presence of a significant proportion of very popular sites among mostly less popular sites. The edge popularity follows a power-law (Pareto) distribution. This means that edge popularity is not comparable to node popularity. Indeed, many edges have a low popularity, whereas few edges have a high popularity.

However, although the distributions are different, we can still observe that edge popularity affects site popularity and vice versa. 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 $R^2 = 0.87$ (network effect) and $R^2 = 0.90$ (site effect) with $p \ll 0.01$, showing 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

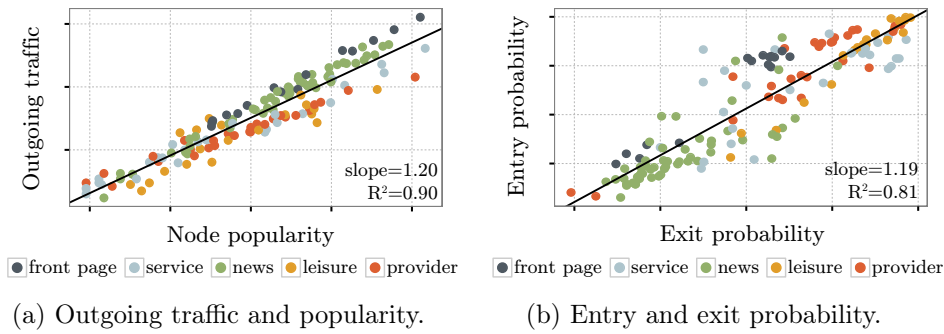


Figure 6.4: Network characteristics. The axes values are removed for confidentiality reasons. The axes of the left plot are in log-scale.

or the network effect is higher. However, we can conclude that there is a *network effect*.

We investigate whether node popularity and its traffic depend on the site category. The correlation between node popularity and outgoing traffic is displayed in Figure 6.4a, the black line represents the regression line. We observe that most of the news sites and front pages are above the regression line, indicating that outgoing traffic increases faster than popularity. In addition, provider sites are below the regression line showing that for these sites the popularity increases faster than the outgoing traffic.

6.4.2 Entry and Exit Points

Finally, we investigate where users enter and leave the network. For each node (site), we calculate the entry and exit probability, that is, the proportion of visits to the site in which the user enters or leaves the network afterwards. Figure 6.4b displays the entry and exit probability per site. We observe that not all sites are equally used to enter or leave the network. There are sites with a high entry and exit probability, and there are sites for which the entry and exit probability is low. We also observe that both probabilities correlate strongly ($R^2 = 0.81$), showing that the entry points are also the exit points in the network. Looking at the site categories, we can see that news sites are less frequently used to enter (leave) the network, whereas leisure sites are often used to enter (leave) the network. Sites with a high entry and exit probability might be even isolated nodes in the network, as users visit solely the site, and thus, enter and leave from there.

Although we could see that there is a strong network and site effect, we could also see that network popularity and traffic can change in a different way. In addition, the node popularity follows a different distribution than the edge popularity. We could also see that the network has specific entry and exit points, that is, some nodes are much more frequently used to enter (leave) the network than others. All these differences reveal that accounting for traffic between sites enhance our understanding of how users engage with a network of sites and the sites itself. Next, we introduce metrics that account for the traffic between sites, and how traffic to and from the network is distributed over the sites.

6.5 Inter-site Engagement Metrics

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. 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 inter-site metrics. When describing them, we specify how they can be used in the context of user engagement.

6.5.1 Network-level Metrics

Network-level metrics are concerned with the online behaviour within the whole provider network. We employ three types of metrics, listed in Table 6.2.

Popularity and activity metrics. The first two types of metrics are an adaption of standard site engagement metrics with a network of sites. They capture the engagement in the provider network, which we refer to as *network engagement*. For this, we adapt the definition of site popularity and activity (see Chapter 4) to our provider networks. The network *popularity* is measured by the number of sessions in which a site in the network was visited (*#Sessions*). The *browsing activity* in the network during an online session is described by the average time users spend in the network (*DwellTimeS*) and the number of sites visited (*#Sites*).

Table 6.2: Network-level metrics: Metrics used to analyse user engagement within a network. $|N|$ refers to the number of nodes in the network.

Metric	Description	Engagement	
		Low	High
Inter-site engagement [IS]			
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
Network engagement			
[POP] #Sessions	Total number of sessions in the network.	0	∞
[ACT] DwellTimeS	Avg. time per session in the network.	0	∞
[ACT] #Sites	Avg. number of sites per session in the network.	0	$ N $

The last type of metrics is concerned with the inter-site engagement in a provider network, that is, they account for the traffic between sites. Section 6.4.1 showed that accounting for the traffic in a network provides new insights compared to network engagement metrics. We define five inter-site metrics which we will introduce next.

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 [266] 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 6.1 we can see that the density of the whole provider network is much lower than the density of the modules (countries) in the network, since 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.

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 [240] 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. Firstly, 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. Secondly, 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. We know from Section 6.4.2 that not all sites are equally used to enter or leave the network. To capture this, we measure how the traffic to and from the network is distributed over the sites, which we call the entry and exit disparity. 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 [83] and adapt it as follows:

$$\frac{\sum_i (g_{max}^* - g_i^*)}{|N| \cdot \sum_i g_i^*}$$

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

6.5.2 Node-level Metrics

Node-level metrics measure the online behaviour on a site within the network. Table 6.3 contains a list of all such metrics used in our work, their description, and their value range.

Popularity, activity and multitasking metrics. We use two metrics that describe the engagement of users with a site (*site engagement*). Following the studies described in Chapter 4 and Chapter 5, we measure the *popularity* of a site by the number of sessions in which the site was visited (*#Sessions*), and the *browsing activity* on a site by the average (median) time users spend on the site during a session (*DwellTimeS*). We additionally employ two *multitasking* metrics defined in Chapter 5: the number of times a site was visited during an online session (*SessVisits*), 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.

We observed in Section 6.4.1 that site popularity and edge traffic differ. Motivated by this observation, 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 the well-known PageRank [202]. 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 navigation of a random user through the network who starts at a certain site. Motivated by [280], we define downstream engagement in the context of traffic networks as follows. We use a discrete-time Markov process to simulate the navigation of a user 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.

Table 6.3: Node-level metrics: Metrics used to analyse user engagement with a site in the network.

Metric	Definition	Engagement	
		Low	High
Inter-site engagement [IS]			
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 and multitasking			
[POP] #Sessions	Total number of sessions on site.	0	∞
[ACT] DwellTimeS	Avg. time per session on site.	0	∞
[MT] SessVisits	Avg. number of visits per session on site.	0	∞
[MT] CumAct	Cumulative activity for the time between visits.	0	∞

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 within 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. In Section 6.4.1, we could see that some sites are more frequently used to enter (leave) the network than others. The two metrics capture these differences by measuring 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.

6.6 Evaluation

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.

In this section, we compare inter-site metrics with site engagement and multitasking metrics. In the following sections, we present two case studies showing how inter-site metrics can be used to enhance our understanding of user engagement. In addition, we performed a sanity check using a network that consists of a selection of sites based in the United States (defined in Chapter 9). We can report that similar results were reached.

We use the daily networks introduced in Section 6.3, namely $YH_{01/08}$, ..., $YH_{31/07}$. We calculate the network-level metrics for each network, and the 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 rho 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 6.4. 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 visit 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 *DwellTimeS*, because the p-value is above 0.01.

Table 6.4: Spearman’s rho between network-level metrics. Correlations with a p-value > 0.01 are not reported (-).

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

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 6.5 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), are also visited frequently when browsing through the network ($\rho(\#Sessions,$

Table 6.5: Spearman’s rho between node-level metrics. Correlations with a p-value>0.01 are not reported (-).

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

$PageRank$) = 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 about the engagement at site level, whereas the page rank relates to sites that are very popular.

6.7 Comparing Provider Networks

The objective is to show that provider networks vary in their inter-site engagement. We compare the five country-based networks (YH_{c1}, \dots, YH_{c5}) with each other. Using network-level metrics we are able to compare engage-

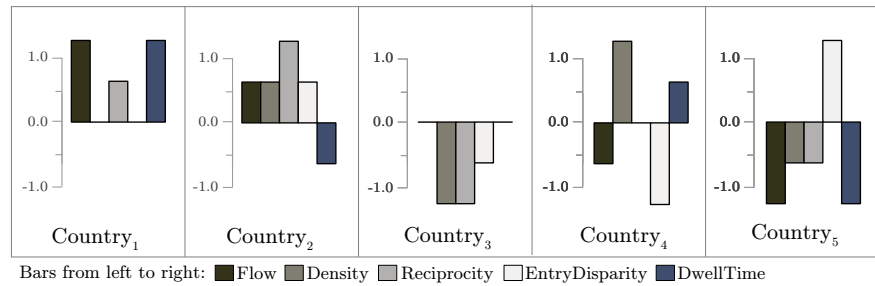


Figure 6.5: Comparing provider networks from different countries using network-level metrics.

ment between networks, as we do it when comparing sites using node-level metrics.

Figure 6.5 depicts the differences between the networks using four selected inter-site metrics. To capture the engagement with the provider network, we employ one network engagement metric: *DwellTimeS*. 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 network engagement can be observed for the first provider network (Country₁). 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 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 (Country₂). 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 *DwellTimeS*). 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 (Country₄). 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 spends 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 (Country₅) 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 *DwellTimeS*).

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 with a network of sites. Indeed, we saw that some networks have a high provider engagement, but a low inter-site engagement, and vice versa.

6.8 Patterns of Inter-site Engagement

We study the different engagement patterns at site-level. Our goal is to show that inter-site metrics provide additional insights compared to those coming from assessing user behaviour within a site. The engagement patterns are determined based on several node-level metrics, *PageRank*, *Downstream*, *EntryProb*, *DwellTimeS*, 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 contains at least 20% of the sites under consideration (155 of them). We use the network constructed of February 2014 (*YH_{Feb}*).

We obtain four clusters, shown in Figure 6.6, 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 the probability difference *PD* as the likelihood that a category occurs in a given cluster with respect to its likelihood that it occurs at all (see Section 3.3 for a detailed definition of *PD*). 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, because 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*).

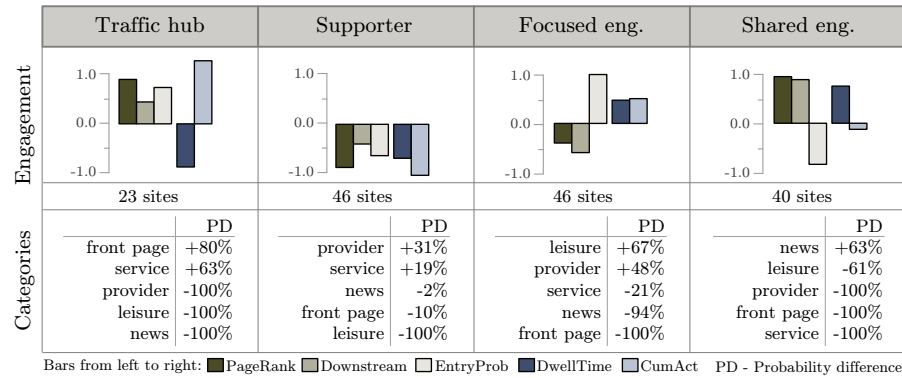


Figure 6.6: Inter-site engagement patterns: (1st row) Site clusters and browsing characteristics. (2nd row) Number of sites per cluster. (3rd row) Likelihood that a category occurs in a given cluster with respect to its likelihood that it occurs at all.

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 *DwellTimeS* 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 *DwellTimeS*), 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 *DwellTimeS*), 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 *DwellTimeS*), 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).

6.9 Discussion

This chapter proposed a methodology to study user engagement in a network of sites. We refer to this type of engagement as *inter-site engagement*. Large internet companies (*e.g.* Amazon, Google, Yahoo) operate a network of sites, offering a variety of services, ranging from shopping to news. We modelled sites (nodes) and user traffic (edges) between them as a network, and employed metrics (at network- and node-level) from the area of complex graph analysis in conjunction with site engagement metrics to study inter-site engagement. This enabled us to consider the user traffic 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 standard 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 employed 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, since linking to them might increase the engagement with the network. We referred to all these metrics as *inter-site* metrics.

We then compared inter-site metrics with site engagement metrics. This brought various insights about inter-site engagement, which would not have been possible with site 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 (*i.e.* users spend a lot of time in the network), but where the inter-site engagement is low (*i.e.* 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.

The chapters in Application II focus on inter-site engagement. Chapter 9 extends the study of the Yahoo provider network by analysing how different dimensions, such as user loyalty and day of the week, affect inter-site engagement, and how links on the sites of a provider network influence the inter-site and site engagement. In Chapter 10, we analyse how users read news across different news providers; that is, we study inter-site engagement in a network of news sites.

Limitations. One limitation of this study is that we did not consider the entire user navigation path, that is the sequence of accessed sites by a user. Our traffic network is a good approximation of how user navigate through the network. However, considering the navigation paths of users might reveal further insights about how users engage with respect to the whole provider network.

In addition, we did not account for sites that do not belong to the provider network but are strongly connected to it (*e.g.* there might be a strong connection between Tumblr and other social media sites). These sites might function as “traffic bridges” between sites of the network (users regularly leave the network to visit these sites but then return to the network afterwards). Accounting for such sites might reveal further insights about inter-site engagement.



PART III

Applications I: Site Engagement

In this part, we perform two case studies that are concerned with user engagement *at* site level. Site engagement and multitasking metrics are employed for the studies.



Native Advertising

The aim of this chapter is to understand how users experience adverts and how the ad quality impact user engagement with the publisher site. We also develop a prediction model to identify high quality ads by analysing their landing pages and relating these to the ad post-click experience.

7.1 Introduction

One of the main source of revenue for services offered via the Web is *online advertising*. Different from traditional *offline* advertising, online advertising uses the Web to deliver promotional marketing messages that are in most cases personalised and tailored to consumers' needs. The two most popular forms of online advertising are *sponsored search* and *display advertising*. Sponsored search shows ads¹ related to queries submitted by users. A display ad embeds a graphic artefact containing a commercial message to deliver to users on a given web page. As opposed to sponsored search, display ads are not delivered in response to any user information need, although they can be personalised based on a profile built on user behaviour.

More recently, *native* (or *in-stream*) advertising has been increasingly emerging as a new form of online advertising designed to specifically offer a user experience that fits with that of the application where the ads are shown. It does so by formatting ads according to the context of the user interface of a given service to make the commercial feel less intrusive. Leading web services have already experienced the promising impact of native advertising,

¹For simplicity, “advertisement” and “advert” are referred to as “ad”.

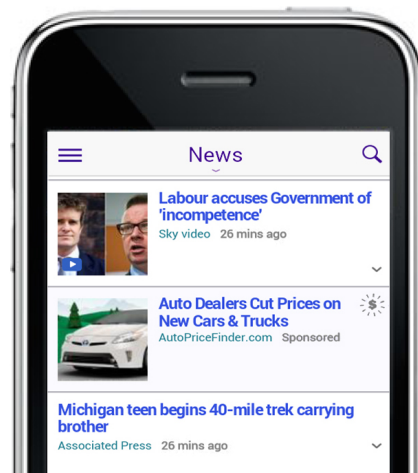


Figure 7.1: Example of a native ad (second item) in a news stream on a mobile device.

for example the launch of sponsored stories on Facebook in 2012 and the release of promoted tweets by Twitter in 2010. Major news sites such as the New York Times, Yahoo News, and the Guardian integrate ads into their streams.² An example of a native ad on a news stream on a mobile device is depicted in Figure 7.1. The ad is the second item in the stream and its appearance is not different to that of the news items in the stream. Therefore, to clearly mark them to users the “Sponsored” label and the dollar symbol “\$” are displayed on the ad item.

Native advertising provides an alternative form of revenue that has the potential to overcome the currently declining publishing business models, which partly arises from the drop in display ad click-through rates.³ Since feed-based layouts, or *streams*, are becoming the predominant interface on mobile devices, native advertising has the appropriate format to serve ads to users [81]. Display ads offer a sub-optimal format, since they require a lot of space on the site and may have a slow loading time.

²<http://america.aljazeera.com/articles/2014/3/8/the-blurred-linesofanativeadvertising.html>

³<http://www.journalism.org/2014/03/26/a-deeper-look-at-the-digital-advertising-landscape>

The aim of native advertising, as well as of other types of online advertising, is to increase publishers' revenue by showing ads that users are likely to click on or to convert (*e.g.* purchasing an item, registering to a mailing list). A user decides if he or she might be interested in the ad content by looking at its *creative*, that is, the ad impression shown within the stream (see Figure 7.1). Therefore it is important that the creative is attractive to users. After a user clicks on the creative he or she is redirected to the ad landing page. How the landing page is experienced by a user, which we refer to as the *ad post-click experience*, is an important factor that can help differentiating between a high quality and a low quality ad. If the landing page is perceived to be of low quality, or if the post-click experience is negative, users might click less on ads in the future, if not stop accessing the service at all with disruptive consequences in the overall number of visitors and therefore in total revenue.

The focus of this chapter is to analyse post-click experience through two well-known site engagement metrics: *dwell time* and *bounce rate*. We look at this in the context of a news stream offered by Yahoo. We demonstrate that the metrics can be used as *proxy* of ad post-click experience for native advertising, and that a negative post-click experiences can have a negative effect on users' engagement with the publisher site (here Yahoo), and the likelihood that the user is clicking on ads again in the future.

As already pointed out, native advertising is especially interesting for mobile devices. However, since existing research focuses mainly on improving advertising on desktop, it is crucial to first understand the differences in ad post-click experience between mobile and desktop. Existing research has shown that user online behaviour differs between the two devices [3; 277], and that, in the context of search, these differences should be considered in the ranking of search results [236].

We therefore compare the post-click experience on mobile and desktop, and show that indeed the experience depends on the device at hand. This also implies that existing models that predict ad post-click experience (*i.e.* ad quality) on desktop may not be applicable on mobile devices as other factors influence user experience with an ad.

We therefore propose a prediction model to identify high quality ads for mobile devices by analysing their landing pages and relating these to the ad post-click experience. We also investigate which landing page features are important for the quality of ads on mobile devices, focusing in particular on features that are unique for the mobile device (*e.g.* mobile optimised landing pages). Finally, we implement and test our model on *Yahoo Gemini*, the new

Yahoo advertising platform for mobile search and native advertising, and we validate its performance by showing the positive impact on click-through rate and post-click experience.

We start by covering previous work in Section 7.2. Section 7.3 analyses how dwell time and bounce rate can be used as a *proxy* for post-click experience. How the post-click experience affects user engagement is studied in Section 7.4. Section 7.5 analyses how post-click experience differs between desktop and mobile devices. The prediction model is introduced and evaluated in Sections 7.6 and 7.7, respectively. The chapter ends with a discussion.

7.2 Related Work

Online advertising. Online advertising has been extensively studied in the context of display advertising (*e.g.* [14; 117; 221]) and sponsored search (*e.g.* [45; 112; 235]). In contrast, to the best of our knowledge, only two studies on native advertising exist. The work of Ritzel *et al.* [218] describes how native advertising can be used by media companies to increase revenue, and Jeong *et al.* [108] proposed a ranking model for the allocation and pricing of ad placements within a stream. We contribute to fill this gap by analysing how users perceive native ads on different devices, and how this affects user engagement with the publisher site.

Users spend an increasing amount of their time online through their mobile phones. This presents unique opportunities for advertisers interested in promoting their products beyond the desktop. Previous works have developed models to predict when to show an ad [198; 205], whereas others have investigated the degree in which mobile advertising is accepted by customers [26] and how users perceive display advertisements on mobile [62]. In this context, studies [81; 218] have highlighted that the dominance of the feed-based structure on mobile makes pop-ups and banners impractical, whereas native ads provide an optimal format as they are seamlessly incorporated to the main feed, thereby avoiding disrupting the overall user experience.

However, mobile online behaviour differs from that of traditional desktop. This should be taken into account when providing native ads on mobile devices. Several works [120; 236], albeit in the context of search, have shown that users use the devices at different times, and that they also search for different content. It has been shown as well that online sessions on mobile differ from that on desktop [3; 277]; users navigate quicker through the

Web when using the mobile device. More importantly, Song *et al.* [236] demonstrated that the differences in user search patterns imply that each device should deploy its own ranking algorithm to provide optimal search results.

Our work is situated in this context; we show that users experience ads differently depending on the device they are using. Motivated by this, we developed a prediction model for native advertising on mobile devices.

Performance measures. Several measures have been proposed to evaluate the “performance” of an ad in terms of the user experience. The most common measure is the ad click-through rate (CTR), which is the number of times the ad was clicked out of the number of times it has been shown (number of ad impressions) [14]. The higher the CTR the better the ad is considered to perform, in terms of attracting the users to click on it.

However, CTR does not account for how users experience the ad when they land on the ad site, namely their *post-click experience*, for which other measures are better suited. For instance, it is possible to measure the probability of users “converting” (*e.g.* purchasing an item, registering to a mailing list) [19; 221]. It is however the case that a positive post-click experience does not necessarily entail a conversion, and conversion rate information is not always available. A less restrictive proxy of post-click experience is the time a user spends on the ad site before returning back to the publisher site, commonly measured through *dwelt time* [278] and *bounce rate* [229]. These measures have been used in online advertising and organic search, *e.g.* to improve the performance of ranking algorithms [128], as well as in recommender systems, *e.g.* to estimate the relevance of an item to a user [277].

In this work, we show that these measures are also good proxies of post-click experience for native advertising, but that they behave differently depending on the device that is used.

Landing pages. The quality of ads is an important aspect to consider, because high quality ads encourage users to click (CTR) and to stay longer on the ad site (post-click experience), which increases revenue. In addition, it has been shown that serving irrelevant or annoying ads has a negative effect on users [31; 91]. Performance measures can be used to evaluate ad quality. It is however the case that this information is very sparse for ads and, in particular, is non-existing for newly inserted ads (“cold-start problem”). Therefore, efforts have been put into building prediction models, where landing page features are used to predict ad quality.

Becker *et al.* [19] showed that conversion rates differ significantly depending on the type of landing page. Also, Choi *et al.* [45] showed that landing pages could be leveraged to better select which ads to return to users in sponsored search. Many efforts have been devoted to categorise landing pages and to use the resulting taxonomies to match ads against search queries [20; 19] (in sponsored search) or web pages [30; 146; 186] where the ad could be shown (in display advertising). For instance, Kae *et al.* [117] focused on automatically categorizing display ad images using image and textual features extracted from the landing page of the ads. Finally, landing page features have been used to predict bounce rate [229], and CTR [217].

Our research adds to this body of work by analysing other features of landing pages for mobile advertising, and how these help to predict user post-click experience measured by dwell time and bounce rate.

7.3 Measuring Post-click Experience

We first define and evaluate two measures of post-click experience. Following from the literature on post-click experience, we use site engagement metrics as our proxy of an ad post-click experience. We define them as follows:

- The average *dwell time* is the average time between users clicking on the ad and returning to the stream.
- The *bounce rate* is the percentage of ad clicks with dwell time, unless otherwise specified, lower or equal to 5 seconds for mobile, and lower or equal to 12 seconds for desktop.

We focus on the measures above, because they are widely used as a proxy of user post-click experience in many contexts, as discussed in Section 7.2.

We randomly sampled 4,000 ads from a large set of native ads served on Yahoo homepage stream in March 2014 both on desktop and mobile. These ads are integrated with the content of the stream and, whilst clearly marked, are designed to look and act just like the stories and format around them. We removed all clicks with a dwell time higher than 10 minutes, since these clicks may correspond to cases when users left the mobile or desktop device and came back later. Doing so removed 1.74% of the total ad clicks. For the bounce rate, the mobile threshold was empirically selected based on the dwell time distribution, which showed a “valley” by around 5 seconds. The desktop threshold was chosen to align with the mobile threshold (*i.e.* by

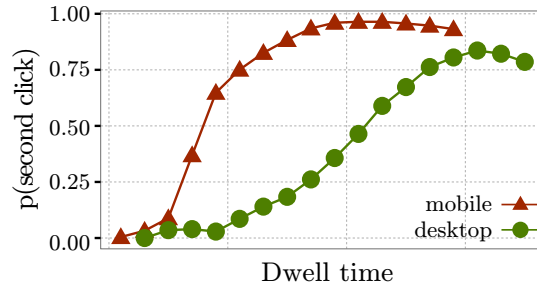


Figure 7.2: Post-click experience. The probability of a second click given dwell time. The x-axis values (log-scale) are removed for confidentiality.

picking the value which corresponds to the same cumulative frequency of “bounce” clicks in both the dwell time distributions). These thresholds fall into the range between 5 and 60 seconds proposed in [159]. Finally, we only consider ads with at least 10 clicks.

Dwell time as proxy of post-click experience. The fact that a user takes time to return to the news stream after clicking on an ad seems a good indicator that the experience is positive: the user browsed the site, maybe converted (*e.g.* purchased a product, registered to the site, shared the page), and finally went back to the stream.

We used a random sample of 200K ad clicks on the mobile stream for which we have records of a click on the ad site. For desktop, we matched the page views associated with the ad clicks to those contained in the Yahoo toolbar’s browsing data, which results in the page views from 30K ad clicks.

The Spearman’s rho coefficient between the number of clicks on the ad site and the ad dwell time is $\rho = 0.65$ for mobile and $\rho = 0.54$ for desktop; the higher the dwell time, the higher the number of clicks on the ad site. Assuming that clicking on the ad site suggests a positive “ad experience”, high dwell time is indicative of a positive post-click experience. The probability of a second click as the percentage of users who clicked on a link on the ad landing page, for given dwell time values, is plotted in Figure 7.2. This probability increases with the time spent *until* the users return to the news stream for both mobile and desktop, further suggesting that dwell time is a good proxy of post-click experience. It is also worth noticing that, to get the same probability of a second click on mobile and on desktop, the dwell time has to be far larger in the latter than in the former. This suggests that dwell time is generally greater on desktop than on mobile.

We recall that dwell time is *not* the time spent on the ad site, but the time between the click on the ad until the user returns to the stream. It can happen that users visit other sites during their session before returning. With the Yahoo toolbar dataset, we are able to verify this. We saw that the higher the dwell time, the higher the probability that users visited other websites. However, for all ad clicks with a dwell time up to 3 minutes, this happens for only 7.4% of the clicks. For dwell time higher than 3 minutes, this percentage increases to 23.3%. Therefore, dwell time is a good proxy of users spending time on the ad site, with longer dwell time suggesting a positive experience with the ad site.

7.4 Effect on User Engagement

Focussing on mobile, we investigate the effect of the ad post-click experience, as measured by dwell time and bounce rate, on long-term user engagement. We divided our dataset into three time-periods, covering a four-week period of user interaction with the mobile stream:

- user *pre-engagement* in a given two-week period;
- user *post-engagement* in the following two-week period;
- user *ad-click-activity* in the last three days of the pre-engagement period and the first three days of the post-engagement period.

Our objective is to compare the pre- and post-engagement periods depending on the ad-click-activity between the two.

We used the *ad-click-activity* dataset to distinguish between a positive ad post-click experience and a negative one. For each ad a , we calculate its mean dwell time $dt_m(a)$ and its standard deviation $dt_{sd}(a)$. Any click c on ad a with $dt_a(c) \leq dt_m(a) - 0.25 \cdot dt_{sd}(a)$ is referred to as a *short click*, and any click c on ad a with $dt_a(c) \geq dt_m(a) + 0.25 \cdot dt_{sd}(a)$ is referred to as a *long click*. Here, $dt_a(c)$ is the dwell time on ad a for click c . These definitions account for the fact that ads differ in terms of their average dwell time. For example, we saw that ads related to beauty products have on average a higher dwell time than those related to finance, simply because the ad experience is different (reading about a product versus registering an interest).

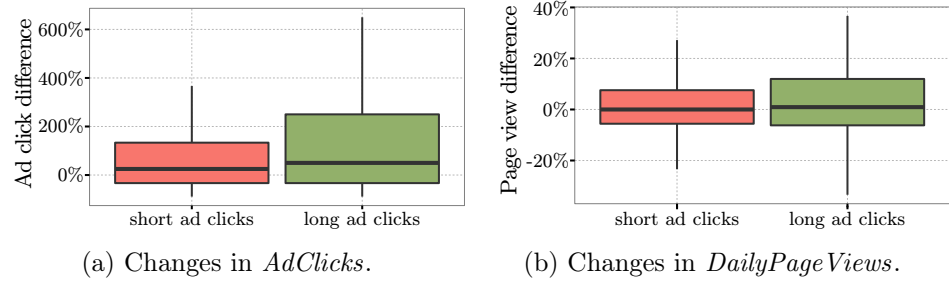


Figure 7.3: Changes in engagement depending on whether the users experienced short or long clicks.

For all users that clicked on at least 3 ads, users with only short clicks were said to have had a negative experience (*shortAdClicker*), whereas those with only long clicks were said to have a positive experience (*longAdClicker*). Having a minimum of 3 clicks allow us to select users that have experienced enough ads to be affected by them. These resulted in two sets of similar size, around 800 users each.

We use two metrics to measure pre- and post-engagement:⁴

- *AdClicks* is the number of ad clicks of a user over the period considered. This metric shows the effect of the ad post-click experience on future ad clicks.
- *DailyPageViews* is the average number of pages a user is viewing within each day over the period considered. This metric shows the effect of the ad post-click experience on future interactions with the stream.

We define the change in the engagement between the pre- and post-engagement time periods as follows:

$$m_{diff} = (m_{post} - m_{pre})/m_{pre}$$

where m is either *AdClicks* or *DailyPageViews*. A value above (below) 0 indicates that post-engagement increased (decreased) compared to pre-engagement, and the extent of the increase (decrease).

⁴Similar results were observed with other metrics and datasets.

Table 7.1: Changes in engagement for short and long ad clickers. We report the average and median (*avg|median*) of the ad clicks and clicks per day difference, and the *p*-values.

	Short ad cl.	Long ad cl.	<i>p</i> -value
<i>AdClicks</i> diff.	90.3% 25.0%	122.6% 50.0%	0.002
<i>DailyPageViews</i> diff.	2.4% 0.0%	5.5% 0.9%	0.000

Figure 7.3 shows the distribution of the two metrics for the two user groups (*shortAdClicker* and *longAdClicker*). We also report the average and the mean of the two metrics in Table 7.1. We use a Kolmogorov-Smirnov test to check whether the distributions are different. The *p*-values are reported in Table 7.1. For both user groups, the ad click activity is increasing. This is likely to reflect that users in both groups are becoming more engaged with the stream and as a result are more likely to click on ads. However, the median *AdClicks_{diff}* for the short ad clicker group is 25.0%; this value is 50.0% for the long ad clicker group. That is, the increase in ad clicks (both in terms of median and average, and distribution) for the long ad clicker group is higher, indicating that a positive ad post-click experience is leading to more ad clicks. The difference is statistically significant (*p*-value < 0.01).

The metric *DailyPageViews_{diff}* has a median close to 0.0% for both user groups. This suggests a similar trend in engagement with the stream, some users becoming more engaged, while others becoming less engaged. Looking at our dataset in more depth, we could identify users getting more engaged with the stream as time passed, and users that were very engaged with the stream. As such their future engagement could not increase further (reaching a certain plateau). However, when looking at the average values and the distributions of *DailyPageViews_{diff}*, we see a larger increase for the long click group, compared to the short click group (5.6% and 2.4%, respectively), suggesting a larger increase in engagement with the stream for users in the positive post-click experience group. The difference, although small compared to *AdClicks_{diff}*, is significant (*p*-value < 0.01).

To conclude, using dwell time to measure the ad post-click experience, we showed that a positive experience has a strong effect on users clicking on ads again, and a small effect on user engagement with the stream. Thus, not only can dwell time be used to measure an ad post-click experience (as a proxy), ensuring that high quality ads are served to users is important for long-term engagement and revenue.

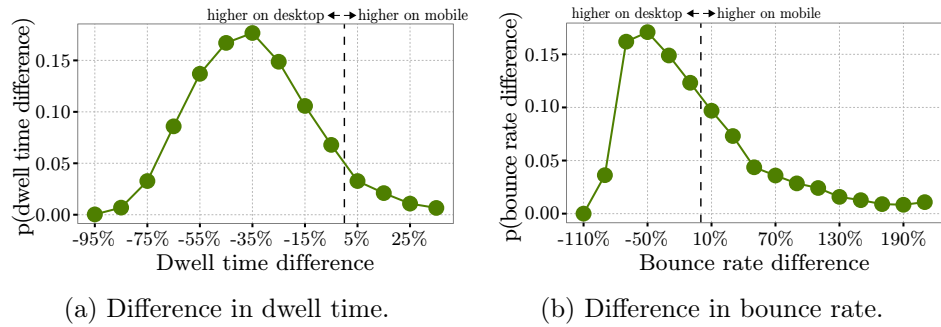


Figure 7.4: Distributions of the differences in post-click experience between mobile and desktop.

7.5 Differences between Mobile and Desktop

We examine the difference in the ad post-click experience, measured by dwell time and bounce rate, between mobile and desktop. First, we compare whether dwell time and bounce rate of an ad differs between the two.

Using Spearman's rho coefficient, for dwell time we obtain $\rho = 0.50$; this value is even smaller for bounce rate with $\rho = 0.23$. Similar correlations were observed when restricting to ads with at least 50 clicks. The correlations are 0.63 and 0.29, respectively. These correlations suggest that users experience ads differently depending on the device they are using.

Next, we calculate for each ad the difference in percentage of their dwell time and bounce rate, when shown on desktop compared to mobile (m refer to dwell time or bounce rate):

$$m_{diff} = (m_{mobile} - m_{desktop})/m_{desktop}$$

The distributions of the percentage differences are plotted in Figure 7.4, (a) for dwell time and (b) for bounce rate. For 92.9% of the ads the dwell time is higher on desktop than on mobile (Figure 7.4a). This is not surprising, as browsing time on mobile has been shown to be shorter generally outside advertising [236]. The highest decrease in dwell time from desktop to mobile is by 35.0%. From Figure 7.4b, we observe that 64.1% of the ads have a higher bounce rate on mobile than on desktop, which is a lower percentage than for dwell time. The highest decrease of bounce rate from desktop to mobile is by 50.0%, which is slightly higher when compared to dwell time. The distribution is skewed to the left however, indicating that many ads



Figure 7.5: Two examples of the same landing page. The one on the left is how it is rendered when the mobile optimisation is not activated. On the righthand side the mobile version.

have a large increase in bounce rate when shown on mobile. For instance, 18.9% of the ads have a bounce rate increase higher than 50.0%.

The low correlation between the ad rankings on mobile and desktop, and the bounce rate differences clearly suggest that the ad post-click experience between mobile and desktop differs. The device has an impact on how users experience the ad implying that a well-performing ad on desktop might have a poor performance on mobile, and vice versa.

Mobile Optimised Landing Pages

We showed that the ad experience on mobile is different to that on desktop. This can partly be explained by the different ways users interact with their desktop and their mobile. However, a preliminary analysis done on the landing pages of the ads in our dataset showed that some of the landing pages were not mobile-optimized, which is likely to have a negative effect on

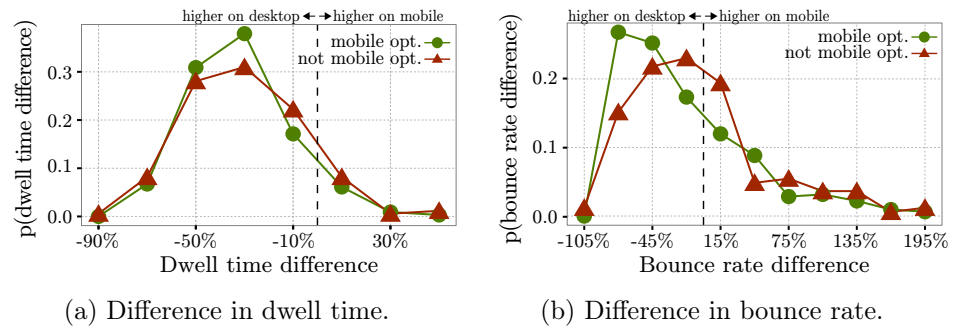


Figure 7.6: Differences in ad post-click experience between mobile and desktop depending on whether the landing page is mobile optimised or not.

users [165]. In Figure 7.5, we show the same landing page rendered on the screen of a typical mobile phone when mobile-specific display options are toggled on and off. In the optimised version, landing pages have typically larger buttons, no long text paragraphs, and a single large image of the product advertised in the middle of the page. To see the effect of this on the ad post-click experience, and the extent to which this reflects the quality of the ad, we designed a mechanism to automatically detect when a landing page is mobile-optimised or not.

We first downloaded the landing pages, rendered them and extracted seven features. We used features such as the size in bytes of the HTML landing page, and whether the page contains an apple touch icon. Next, we manually labelled a sample of 259 ads as mobile-optimised (**Opt**) or non-optimised (**Npt**). Our training set consisted of 108 **Opt** and 151 **Npt** ads. We fed the feature representation of the landing pages in our annotated dataset to a Gradient Boosted Decision Tree classifier, and we estimated its quality using leave-one-out cross validation (LOOCV). The classifier reached an F_1 score of 0.995, demonstrating its high accuracy.

We then tested whether dwell time and bounce rate of a landing page correlates with its property of being **Opt**. Using a sample dataset of 500 ads, with approximately 65.0% of them being mobile-optimised and 35.0% non-mobile-optimised, we conducted a similar experiment to Section 7.5, but with **Opt** ads and **Npt** as our classes. The results are shown in Figure 7.6. We employ a Kolmogorov-Smirnov test to verify whether the differences are significant.

The distribution of the dwell time difference is very similar for both groups, **Opt** and **Npt** (Figure 7.6a). The average dwell time decreases by 31.8% (*median* = 31.8%) for **Opt** landing pages, and by 28.9% (*median* = 33.6%) for **Npt** landing pages. The difference is not significant (p -value = 0.20). Whether a landing page is optimised does not influence how long users spend on the mobile ad site.

When considering bounce rates, however, we observe differences (Figure 7.6b). The average bounce rate decreases by 6.9% (*median* decreases by 30.4%) for **Opt** landing pages but increases by 13.4% (*median* decreases by 11.5%) for **Npt** landing pages. These differences are statistically significant (p -value = 0.003). Therefore, mobile-optimised landing pages have a positive influence on users, as they are less likely to lead to bounce, relative to when shown on the desktop. Studies looking at ad post-click experience in the mobile context should account for this property of the landing page, which by itself is not surprising. However, since we observe no difference in dwell time other features of the landing pages influence user ad post-click experience. This is the motivation for the landing page analysis discussed next.

To summarise, the results show that post-click experience differs between mobile and desktop making it necessary to evaluate ad performance on desktop and mobile separately. In addition, we have shown that mobile-optimised landing pages can have a positive effect on post-click experience, but also other factors seem to influence how users experience ads.

7.6 Predicting High Quality Ads

In the previous sections we studied how dwell time, bounce rate, and ad post-click experience relate to each others. In this section our goal is to devise a method to predict the “quality of the ad” as perceived by the user, with the idea that a high quality ad leads to a positive experience with the ad landing page. This task is different from that of sponsored search. Whereas in the latter, the user’s information need can be (partially) inferred by exploiting the information conveyed by the user query, in the case of native ads, as well as that of display advertising [117], this information is not available. There are no actual signals relative to what a high quality ad means to each specific user with respect to its relevance to that user. The only quality indicators are the click-through rates (users clicking on the ad), the dwell time and bounce rate (users spending time on the ad).

The previous sections demonstrated dwell time to be a good proxy of user post-click experience and bounce rate to provide additional insights about user post-click experience. In this section, we use these metrics to identify high quality native ads on the basis of their landing pages, as experienced by mobile users.

7.6.1 Problem Definition and Methods

Native ads ranking is conceptually the same as ads ranking in sponsored search. Let $P_{\text{click}}^i \in [0, 1]$ be the predicted probability of an ad i being clicked, and let $\text{bid}^i \in \mathbb{R}$ be the amount of money the advertiser is willing to pay for its ad to be shown. Ranking the ads is done by computing the *expected cost per click* $\text{eCPC}^i = P_{\text{click}}^i \cdot \text{bid}^i$ for each ad, and later sorting them in descending order of this value.

Traditionally, the focus in sponsored search has been to build sophisticated models to predict P_{click}^i . Our aim is to predict P_{SAT}^i which is the *conditional probability* of a user being satisfied *given* that he or she clicked on an ad i . The goal is thus to estimate the overall *joint probability* of clicking on an ad *and* being satisfied by its landing page. More formally, we want to compute $P_{\text{HQ}}^i = P_{\text{click}}^i \cdot P_{SAT}^i$.⁵ Finally, the ranking of ads will be computed as $\text{eCPC}_{\text{HQ}}^i = P_{\text{HQ}}^i \cdot \text{bid}^i$.

We thus want to predict a class label $Y_i \in \{-1, 1\}$ for a given landing page X_i identified by a feature vector $\phi(X_i)$. The class label Y_i is 1 if X_i is a high quality page, -1 otherwise. Concretely, we aim at estimating the following probability density function

$$P(Y_i = -1 | \phi(X_i)) = 1 - P(Y_i = 1 | \phi(X_i))$$

that, in practice, corresponds to the joint probability P_{HQ}^i .

Three definitions of high quality ads. We consider three definitions of *high quality* ads, based on dwell time and bounce rate on their landing pages. Formally, with a web page X_i accessed by a set of n_{X_i} users we associate two real numbers, $\delta_{X_i} > 0$ its mean dwell time computed over the set of n_{X_i} users, and $\beta_{X_i} \in [0, 1]$ its bounce rate computed as the fraction of the n_{X_i} leaving X_i before a given time threshold.

⁵Note that P_{SAT}^i is conditioned on P_{click}^i .

1. *High Dwell Time.* $Y_i = 1$ when $\delta_{X_i} > t_\delta$ where t_δ is a threshold defined on dwell time. Under this assumption $P_{\text{HDT}}^i = P(Y_i = 1 | \phi(X_i)) = P(\delta_{X_i} > t_\delta)$.
2. *Low Bounce Rate.* $Y_i = 1$ when $\beta_{X_i} < \tau_\beta$ where $0 < \tau_\beta < 1$ is a bounce rate threshold value. Under this assumption $P_{\text{LBR}}^i = P(Y_i = 1 | \phi(X_i)) = P(\beta_{X_i} < \tau_\beta)$.
3. *CombHQ.* $Y_i = 1$ when $\delta_{X_i} > t_\delta$ and $\beta_{X_i} < \tau_\beta$ where τ_β and t_δ are defined as before. We want a high quality landing page to such that a user enters, does not bounce, and stays long enough to interact. Thus $P_{\text{CMB}}^i = P(Y_i = 1 | \phi(X_i)) = P(\beta_{X_i} < \tau_\beta, \delta_{X_i} > t_\delta) = P(\beta_{X_i} < \tau_\beta) \cdot P(\delta_{X_i} > t_\delta)$, where we assume the two events, low bounce rate and high dwell time, to be disjoint.

Ad landing page features. Inspired by previous work [19; 20; 45; 117] exploiting features extracted from the landing pages to categorise ads, we defined three sets of features: CONT considers the content characteristics of the landing page; SIM considers the similarity between the ad creative as displayed in the stream and the content of the landing page; and HIST considers the history of the ad performance. A list of all features is given in the Appendix in Section A.1.1. While aforementioned works aim apply to ads served on desktop, our goal is to use ad landing page features to discriminate between high quality and low quality native ads shown on mobile streams.

Prediction quality. Using the features defined above we trained a model to predict P_{HQ}^i using different learning methods. We used three well known methods: logistic regression [281], Support Vector Machines (SVM) [50], and Gradient Boosted Decision Trees (GBDT) [85]. We used the implementations of these methods available in the Python `scikit-learn` package.⁶ The probability values were extracted using the implementation available with this framework. We adopted standard parameters for each method. For logistic regression we set C , the inverse of regularisation strength, to 100, and $L1$ as penalty norm, and 0.01 as the tolerance value for stopping the optimisation. For the SVM classifier we adopted a *RBF* kernel with a penalty parameter C of the error term equal to 1.0, 0.0 as the gamma kernel coefficient, and 10^{-3} as the tolerance used in the stopping criterium. Finally, for the GBDT classifier we generated a forest of 100 trees with a max depth of each tree of 4, and a learning rate of 0.01.

⁶<http://scikit-learn.org>

Table 7.2: Prediction performance on models built on ads data from March 2014 and tested on April 2014.

Features	Method	t_δ	τ_β	AUC	F ₁	MCC
Dwell time						
C	svm	40	-	0.83	0.79	0.67
C-S	svm	40	-	0.83	0.79	0.67
C-H	svm	40	-	0.83	0.80	0.68
C-S-H	svm	40	-	0.83	0.80	0.68
Bounce rate						
C	logistic	-	0.22	0.61	0.74	0.27
C-S	logistic	-	0.22	0.61	0.72	0.23
C-H	logistic	-	0.22	0.86	0.86	0.71
C-R-H	logistic	-	0.22	0.85	0.86	0.70
CombHQ						
C	svm	50	0.2	0.79	0.67	0.61
C-S	svm	50	0.2	0.79	0.67	0.61
C-H	svm	50	0.2	0.81	0.68	0.62
C-R-H	svm	50	0.2	0.81	0.68	0.62

7.6.2 Offline Models Evaluation

To assess the validity of our prediction models, we ran a traditional offline evaluation based on historical data.

Experimental setup. We used the three different definitions of high quality landing pages. For each, we report the performance of predictors using the three standard metrics, Area Under the ROC Curve (AUC), F₁, and the Matthews Correlation Coefficient (MCC) [171]. The latter is a correlation measure between predictions and labels taking into account the popularity of each class. The dataset used to run the experiments is a uniformly generated sample of our ad set. As a training set we extracted a sample of 1,500 ads shown in March 2014 to users of the system. The test set contains a sample of 550 ads shown in April 2014. In all the tests we conducted we experimented with several thresholds for dwell time (t_δ) and bounce rate (τ_β). Finally, we tested several combinations of features: content-based or C features; similarity-based or S features; and history-based or H features.

Table 7.2 reports the best results for predicting the probability of high dwell time, low bounce rate, and the combination of them. A detailed overview of the results can be found in the Appendix in Section A.1.2. First, the various classification methods perform similarly (with a slight advantage of

Table 7.3: Top-15 ranked features using random forest with two sets of features: C, and C-S.

Rank	C	C-S
1.	[C] <i>clickToCall</i> (0.331)	[C] <i>clickToCall</i> (0.313)
2.	[C] <i>windowSize</i> (0.125)	[C] <i>summarisabilityScore</i> (0.109)
3.	[C] <i>numClickable</i> (0.113)	[C] <i>numConceptAnnotation</i> (0.086)
4.	[C] <i>numInputRadio</i> (0.089)	[C] <i>numInputRadio</i> (0.086)
5.	[C] <i>numDropdown</i> (0.084)	[C] <i>numDropdown</i> (0.077)
6.	[C] <i>numImages</i> (0.064)	[C] <i>numClickable</i> (0.075)
7.	[C] <i>numInputCheckbox</i> (0.031)	[C] <i>numImages</i> (0.048)
8.	[C] <i>imageHeight</i> (0.029)	[C] <i>windowSize</i> (0.036)
9.	[C] <i>nounsSumOfScores</i> (0.026)	[C] <i>imageHeight</i> (0.029)
10.	[C] <i>numConceptAnnotation</i> (0.024)	[C] <i>nounsSumOfScores</i> (0.028)
11.	[C] <i>viewPort</i> (0.021)	[S] <i>similarityNoun</i> (0.023)
12.	[C] <i>media</i> (0.021)	[C] <i>numInputCheckbox</i> (0.023)
13.	[C] <i>tokenCount</i> (0.014)	[C] <i>viewPort</i> (0.016)
14.	[C] <i>numInputString</i> (0.011)	[C] <i>isMobileOptimised</i> (0.014)
15.	[C] <i>imageWidth</i> (0.010)	[C] <i>media</i> (0.012)

SVM over the others)⁷ over all the metrics we tested. Similarity-based features perform bad as they never increase (and in some case are detrimental to) any of the experimented metrics. Finally, history-based features are very important as they boost, for instance, AUC above 0.8 when combined with content. Content-based features alone are already achieving (with the exception of CombHQ classifiers) high values with all three metrics.

It is worth remarking that history-based features are very sparse for ads and, in particular, are non-existing for newly inserted commercials. Content-only classifiers, thus, can always be used, because they provide the perfect solution to the “item cold-start problem” that will be experienced with many ads. The high quality classifier does not reach high performances as, in our opinion, the amount of true positive instances is relatively small when compared to the other cases. Finally, the choice of the threshold has some effect.

7.6.3 Feature Ranking

For information, we show the importance of the two sets of features that can be used for the “item cold-start problem”, namely C (content characteristics), and S (similarity between ad creative and content). We use the technique proposed in [28] implemented in the `scikit-learn` toolkit. This technique uses *Random Forest*, which can be used to induce a ranking of the “importance” of

⁷For high dwell time and high quality metrics.

features in a regression or classification problem. Table 7.3 shows the top-15 features ranked according to their importance scores as output by random forest. Each column refers to the ranking of a specific set of features. The i -th row contains the i -th ranked feature, along with its category (C, or S) and importance score, for each set.

When the classifier is trained using only content features (C) or both, content and similarity features (C-S) *clickToCall* is the most important signal. The first similarity feature (*similarityNoun*) is ranked 11-th when using the C-S set. This means that similarity features do not provide significant insights to discriminate between high and low quality ads. We also observe that features related to the functionality of the landing page (*e.g. numDropdown, numInputRadio*) are more important than features related to the content quality (*e.g. tokenCount, nounsSumOfScores*), and the aesthetic appeal (*e.g. isMobileOptimised, media*). Only when considering the similarity features (C-S) as well, two of the content-quality-related features become more important (*summarisabilityScore* and *numConceptAnnotation*).

Using logistic regression, we can see that the functional features all have negative coefficients implying that they are used from the model to identify low quality ads. We speculate that these features are most likely embedded in a form on the landing page, and as such prevents users to continue browsing through the site, since it is not user-friendly to fill out a form on a mobile device, or users are not willing to share private information. However, it is also possible that such forms point to a specific type of advertisements such as insurances and loans, which ends up not being interesting for the users.

7.7 Online Bucketing Evaluation

To measure the impact that such an ad ranking scorer has on users we conducted an *online* evaluation. We implemented and tested our ad ranking scorer on *Yahoo Gemini*, the new Yahoo advertising platform, and we validate its performance on the mobile news stream app running on iOS.

Prediction model. In the previous section, we showed that dwell time is a good proxy of an ad post-click experience. The logistic regression model showed very good performance in predicting high dwell time, *i.e.* dwell time being above a given threshold. Although the SVM performances were sometimes slightly higher, a logistic regression model supports quick update operations, which is important when deployed in production. We therefore decided to deploy such a model using only content-based features to allow

full coverage of the ads in the database, and not just those for which we have historical data. We use 40 seconds as our threshold, as in our dataset, it corresponds to the median of the overall dwell times distribution. We also chose to deploy the version predicting high dwell time since we wanted to serve ads on which users spend time.

Experimental setup. We split the incoming traffic into two *buckets*, *i.e.* *baseline* and *ad quality*. In the first bucket, ads are served using the existing ranking scheme, *i.e.* the expected cost per click, whereas in the second bucket ads are served according to the newly proposed ranking scorer that accounts for the ad post-click experience (*i.e.* the probability that users do not return to the stream within the next 40 seconds).

We measure user’s post-click experience with dwell time and bounce rate. Specifically, we compute the *median* of the former to deal with the high variance of dwell times in the two buckets. For the bounce rate we report, instead, the *average* of the values. Bounce rate is already a normalised score in the $[0, 1]$ range and also exhibits small variance in both buckets. In fact, it makes no sense to show the median for bounce rate given that, due to the shape of the underlying distribution of short clicks, the difference in the median values will be extremely high, therefore unrealistically favouring the *ad quality* bucket.

We considered two distinct datasets of ad clicks, randomly sampled from May to June 2014. The first dataset is drawn from the baseline bucket and contains only clicks on ads ranked by the baseline scorer. The second dataset is drawn from the ad quality bucket and contains all the clicks of ads served by the ad quality ranking scorer.

We conduct three analyses, at the (*ad-*)*click-level*, at the *ad-level*, and at the *user-level*. In all of three, we evaluate the two buckets performance. First, we compare the performance accounting for all ads (users) as they appear in the two datasets; we refer to this experimental setting as **All**. Then, we measure the performance limited to only those ads (users) common to the two datasets; we call this **Shared**. Finally, we focus only on those ads (users) that appear in only one of the two datasets; we refer to this as **Unique**.

Click-level analysis. We discuss how the daily click-through rate behaves on the two buckets. When assessing the effect of any change, *e.g.* in a ranking algorithm, it is important to do so over a long period of time, because an increased performance shortly after the change may not translate in the long run to better user experience [130].

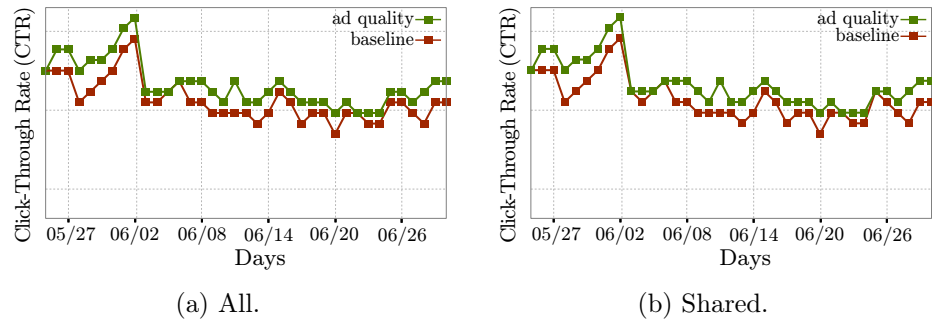


Figure 7.7: Daily Click-Through Rate (CTR): *baseline* vs. *ad quality*.

We collected around 14,500 ads from the dataset relating to the baseline bucket, and about 12,500 ads from the ad quality bucket (All). More than 11,000 ads are shared between the two buckets (Shared), and therefore can be considered high quality ads.

Looking at these two settings (All and Shared) we can see that the daily click-through rate is always higher for the ad quality bucket, as shown in Figure 7.7. This means that the probability of a user clicking on an ad increases in the ad quality bucket. Interestingly, the two time series are perfectly correlated (Pearson’s $r = 1.0$). The *paired t-test* applied to each pair of time series samples shows that the differences between the two samples, though correlated, are statistically significant ($p = 0.01$).

If we consider only the (high quality) ads shared between the two buckets the chance of clicking on an ad seems to depend on which bucket the ad is served. Intuitively, if a high quality ad is served together with other, lower quality ads (*i.e.* baseline bucket), users may not perceive it as valuable as it is, and thus the probability of clicking on it decreases.

However, if the *same* high quality ad is served with other high quality ads (*i.e.* ad quality bucket) the users may be more likely to click on it because they have been exposed to ads leading to positive experience. Therefore we conjecture that the perception of the ad quality is influenced by the ads to which a user has been “exposed” and by how they have been experienced (*i.e.* either satisfactory or not), which is what Figure 7.7b suggests.

Ad-level analysis. The second analysis shows how dwell time and bounce rate behave on the two buckets, from an “ad perspective”. We first remove all the ad clicks with dwell time greater than 10 minutes (following from

Table 7.4: Differences (%) in dwell time and bounce rate between *ad quality* and *baseline* ads, and *ad quality* and *baseline* users.

	Ad-level			User-level		
	All	Shared	Unique	All	Shared	Unique
Dwell Time (<i>median</i>)	+30.00%	+20.00%	+35.71%	+25.00	+20.00	+30.00
Bounce Rate (<i>avg</i>)	-6.67%	0.00%	-25.00%	-12.00	-12.50	-12.00

Section 7.3). Afterwards, we consider only those ads that received at least 10 clicks. This is done to avoid the effect of outliers and ensure that we have enough clicks to calculate bounce rates.

Finally, a click is considered a bounce if its dwell time is less than or equal to 5 seconds. This resulted in around 1,000 ads in the baseline bucket and 700 in the ad quality bucket (**All**), with around 600 ads common to both buckets (**Shared**). We should note the ads used in this analysis form a subset of those used in the first click-level analysis. Table 7.4 shows the relative differences (in percentage) of dwell time and bounce rate as computed from the two buckets.

Unique is the dataset comprised of ads that are only in one of the two buckets. Each cell of the table refers to the relative difference (in percentage) between the statistics as computed from the ad quality and baseline buckets, respectively. Note also that while a positive difference is desirable in the case of dwell time (*i.e.* showing the ad quality bucket exhibits more time spent on the ad landing page), for the bounce rate we aim at reducing the number of short clicks. That is why we prefer a negative difference.

In all the experimental settings, the ad quality bucket outperforms the baseline bucket. This is particularly visible when considering **Unique** ads. Interestingly, when looking at the **Shared** ads, the median dwell time is still higher when the ads are served as part of the ad quality bucket, compared to the baseline bucket. In Figure 7.7b, we saw that the click-through rate for the **Shared** ads is higher in the ad quality bucket. Therefore, not only more users are attracted to the ads when served within the ad quality bucket, these additional clicks also lead to a positive post-click experience.

Similarly, the average bounce rate is lower with the ad quality bucket, which implies a lower probability of bouncing back once users click on ads that have been deemed to be of high quality. Interestingly, for **Shared** ads, there is no difference in bounce rate. This suggests that the quality of an ad is a

characteristic of the ad itself (*i.e.* its landing page as experienced by users), and does not depend on what other ads are served within the same session.

We compare the distribution of dwell time and bounce rate rate, as observed in the two buckets. We run a Kolmogorov-Smirnov test for two samples on each pair of observations to test if both samples might have been drawn from the same underlying probability distribution. For all cases except one, this is not the case with p -value lower than 0.01. The exception is with the bounce rate on the **Shared** setting. This however complies with the results shown in Table 7.4, where no difference exists between the two buckets for the average bounce rate for the **Shared** setting, further confirming that the quality of an ad comes from “itself”.

User-level analysis. The aim of the last analysis is close to that of the previous one yet from a “user perspective”. We remove all the ad clicks having dwell time larger than 10 minutes and a click is considered a bounce as long as its dwell time is at most 5 seconds. Furthermore, we take into account only those users who clicked on at least 2 unique ads. This resulted into around 16,000 users in the baseline bucket and 11,000 in the ad quality bucket, with about 2,700 individuals shared between the two buckets. The results are reported in Table 7.4.

The median dwell time is higher for the ad quality bucket, which means that when users are served ads deemed of high quality, they spend time on the ad landing page before returning to the stream. In particular, looking at the **Shared** setting (*i.e.* users appearing in both buckets), serving high quality ads indeed promotes a positive post-click experience. This is further accentuated when users experience only high quality ads (as seen with **Unique**).

Concerning the average bounce rate, this is computed as the average fraction of bounce clicks for each user. This is different from the actual definition of bounce rate, which instead is formulated at the ad-level. Still, we observe a decrease of the average bounce rate in the ad quality bucket. This relates to the fact that more high quality ads are served in the ad quality bucket, which leads to fewer users bouncing back after clicking on them. Finally, the difference between dwell time and bounce rate distributions is statistically significant ($p \ll 0.01$) using Kolmogorov-Smirnov test.

Overall this section shows that returning high quality ads is important. Not only this increases CTR, and as a likely consequence revenue in the long-term, it has a positive effect on users, as seen by the increase in dwell time and decrease in bounce rate.

7.8 Discussion

This chapter provided new insights about how users perceive content (ad) quality and its effect on user engagement. We looked at this in the context of *native advertising* which is a new form of online advertising that allows advertisers to display their commercial contents integrated into a publisher's main stream.

We first showed that dwell time and bounce rate (two site engagement metrics) are appropriate proxies of ad quality, with respect to how ad landing pages are experienced by users (the post-click experience). We then investigated the effect of ad quality on long-term user engagement, and showed that a poor post-click experience results in users clicking less on ads again, which leads to a loss in revenue. It might also cause users visiting the service less or stop visiting it at all.

In conclusion, serving high quality ads is essential for publisher's revenue. It encourages users to click on ads and maybe to convert, but it also keeps users visiting the service and therefore to continue clicking on ads.

We also compared the post-click experience on mobile and desktop, and showed that users experience ads differently depending on the device they are using, that is, their perception of quality is device-dependent. This might be caused by the fact that the mobile user interface is much smaller and that the interaction possibilities are much more limited (*e.g.* using a keyboard is more efficient than using a touchpad). However, the fact that users use the devices for different information needs could also be a reason for these differences. Following [236], ads related to celebrities and movies might perform much better on mobile than ads related to loans and insurances (finance).

Motivated by this observation, we put forward an approach that analyses ad landing pages, and shows how these can affect dwell time and bounce rate. We then developed a prediction model for ad quality based on dwell time deployed on *Yahoo Gemini*, the new Yahoo native advertising platform.

We carried out both, offline and online evaluation, to assess the validity of our solution. Results of the offline evaluation showed that, using dwell time as a proxy of post-click experience, we could deliver ads that were of higher quality. The online evaluation revealed that our solution positively affects the post-click experience and click-through rate, which increased when high quality ads were served.

Limitations. There are several landing page features that we did not consider in our prediction model such as readability, pagerank, sentimentality level of the landing pages. We are also currently carrying out user studies to understand how users perceive the quality of the landing pages, and how this can be translated into additional features.

In addition, we did not look at the impact of the ad placement within the content stream as well as the relevance of the ad to the surrounding context. Finally, we only focused on two metrics to measure post-click experience. Further exploring how users interact with the ad site (clicks, mouse movements) might enhance our understanding of post-click experience and ad quality.



Reader Engagement in Wikipedia

This chapter studies how readers engage with Wikipedia by characterising their reading preferences and behaviours, and illustrates how reader engagement can provide valuable insights to Wikipedia's editor community.

8.1 Introduction

Peer-production communities have transformed the way people use and experience the Web. The collective action of these communities usually evolves around a digital artifact, such as an online encyclopedia or a piece of software. Wikipedia is a famous example of a peer-production community, and is the focus of our study.

Wikipedia is a multilingual, web-based, free encyclopedia, written collaboratively by a large number of volunteers. Since its creation in 2001, Wikipedia has grown into one of the most visited websites, attracting 530 million unique visitors monthly (October 2013).¹ As of November 2013, Wikipedia was available in 287 languages and comprised about 30 million articles. The English Wikipedia, the largest language version, had more than 30,653 active contributors working on over 4.5 million articles.² It was ranked as the

¹<http://reportcard.wmflabs.org/>

²Registered (and signed in) users who made 5 or more edits in a month.

8th most popular website on the Internet in the US, where popularity is measured by the number of page views.³

Scholars have attributed the success of Wikipedia to its *production side*, that is the quality of its articles and authors' participation [107; 129; 242]. Thus, Wikipedia's production side has been the focus of numerous studies while another group of Wikipedia users, the readers, has not been much studied. A literature review by Okoli *et al.* [199] covering 477 research studies on Wikipedia showed that 42% of the studies mostly centered on issues related to participation, *i.e.* how editors⁴ create and edit articles, resolve disputes, or organise their community. Only 20% of the studies are related to readers, the *usage side* of Wikipedia, such as examining the popularity of articles or topics in Wikipedia. Less than 1% of the reviewed studies looked at users' reading preferences and only one study investigated reading behaviour [199].

One reason for the limited focus on Wikipedia readers might be how scholars consider the role of passive users, *i.e.* the readers, in online communities. Readers are often considered to not provide any visible contribution to the community, and have been referred to as "lurkers" or "free-riders" who are "more resource-taking than value-adding" [132]. When scholars showed interests in this user group, it was mostly because reading is often seen as the prerequisite for becoming a contributor [144; 210; 211]. For example, Halfaker *et al.* [95] carried out several experiments to encourage Wikipedia readers to become contributors.

An exception is the work by Antin *et al.* [8], who claim that reading can be seen as a form of participation and is therefore valuable: the fact that a user is reading an article and not editing could be interpreted as an indication of an article's quality, such as its reliability [2]. Thus, reading activity – the usage side – can provide valuable insights to editors – the production side.

Other peer-production communities, such as open source software development projects, have included the usage side into their definition of success. They use measures that typically revolve around quantifications of volume related to the number of accesses to a particular project's product or outcome [53; 109].

Inspired by these perspectives, we conjecture that the same paradigm can be used in the context of Wikipedia. Instead of looking exclusively at the

³<http://www.comscore.com/Insights> (for desktop access). The ranking is 9th when accounting for mobile access.

⁴In this chapter, we use author and editor interchangeably.

production side (the editors), we analyze the usage side (the readers) and discuss how our analysis can inform Wikipedia’s production side. In other words, this work makes readers and its activities *visible* and *valuable* to Wikipedia’s editor community. Using site engagement and multitasking metrics defined in Chapter 4 and Chapter 5, we explore how readers engage with Wikipedia by studying their *reading preferences and behaviours*. We then compare readers’ *engagement* with Wikipedians’ *editing activity*.

We show that the most read articles do not necessarily correspond to those frequently edited, suggesting some degree of non-alignment between user reading preferences and author editing preferences. We also show that popular and often edited articles are read according to four main patterns, and that how an article is read may change over time. Although we observe that readers’ engagement is mainly driven by their interests, we assume that for promoting a successful reading experience the quality of the articles is important as well. We therefore demonstrate, through examples, how readers can provide valuable insights to Wikipedia’s editor community and that they are *not* resource-taking *but* value-adding.

First, we review existing literature on reading preference and reading behaviour and show that current knowledge is limited and rather exploratory. Section 8.3 introduces the datasets used in this chapter. Users’ reading preferences and behaviours are analysed in Sections 8.4 and 8.5, respectively. Section 8.6 discusses how the findings can be used from Wikipedia’s production side for their editorial work.

8.2 Related Work

Few studies about reading preference of users on Wikipedia exist. Sporerri [238] examined readers’ interests with respect to the topics they read about. The analysis, based on view count, showed that the most accessed articles were in the areas of entertainment (music, films, TV series), politics/history (politicians such as George W. Bush, historical events such as World War II), and geography (places such as Paris or countries such as USA). This aligns with the study reported by Waller [260], who investigated search queries from Australians to Wikipedia. In general, people are more interested in “lighter” topics such as entertainment than in more “serious” or advanced topics. In this work, we also show that readers in English Wikipedia have similar interests. However, a survey carried out on university students regarding the specific websites they have in mind when searching

for information, reveals that 34% of the students would use Wikipedia for factual information and only 6% indicated thinking about Wikipedia when searching for entertainment related information [246].

Preference for both information searching and entertainment is a shared characteristic of Wikipedia readers and editors according to West *et al.* [272]. The authors leveraged data from a browser toolbar to investigate differences in web usage between Wikipedia editors, readers and Internet users who did not access Wikipedia. They found that editors are “information-hungry” and “entertainment-loving”, as they spend more time on news and search, but also on YouTube and other entertainment sites; Wikipedia readers’ preferences are in a middle ground between those of editors and users not accessing Wikipedia.

A comparison of reading behaviour and editing activity in Wikipedia was performed by Reinoso *et al.* [216]. The authors compared for different language editions of Wikipedia the number of page views and the number of edits performed on them. For languages such as English, German and Spanish, the number of views and edits were highly correlated. This was not the case for Japanese and Dutch.

Reading behaviour has been studied by Ratkiewicz *et al.* [214], who explored the dynamics of the popularity of Wikipedia topics. Popularity was defined as the number of hyperlinks linking to an article and the number of clicks to it. The authors found that almost all articles experience a burst just after their creation and the majority of articles receive little attention thereafter. Only few articles show intermittent bursts later in their lifetime. Ten Thij *et al.* [248] built a model to explain bursts in reading behaviour caused by featuring an article on Wikipedia’s main page.

Finally, two studies looked at how readers navigate within Wikipedia. Helic [99] analyzed users’ click paths on Wikigame, where users must find the way (clicking links) from one randomly selected Wikipedia article to another. The author showed that users are very efficient at navigating; indeed users easily found short paths between the randomly selected articles. Gyllstrom *et al.* [94] investigated different browsing patterns on Wikipedia. They found out that user online behaviour depends more on the page topic than on the linking structure. They suggested that understanding different browsing strategies can help editors to better present or organise their content.

These studies demonstrate the still limited knowledge on user reading behaviour on Wikipedia, in particular in relation to Wikipedia’s production side. In the following sections, we carry out an analysis to gain insights about reading preference and behaviour in Wikipedia, and discuss how these insights can add value to Wikipedia’s peer-production side. We start by describing the datasets used in our work.

8.3 Datasets

Our study is based on data collected over a period of 13 months (September 2011 to September 2012) from various sources for the English Wikipedia. In the first part of our analysis (Section 8.4.1), we use all Wikipedia articles to determine and study the most popular topics. To work on a more homogeneous dataset and avoid the effect of structural differences between different types of articles, we then focus for the rest of our analyses on a specific sub-set of articles – namely biography articles which contain descriptions of persons, such as actors, singers and historical figures. Biography articles form the most popular topic in Wikipedia. This approach was already followed in previous research [80]. To detect biographies, we considered all articles belonging to the Wikipedia category “Living people”, as well as to the categories “Births by year” and “Deaths by year” and recursively to their subcategories. We then removed categories that did not contain biographies, and articles that were lists of biographies.

Page view data. As a measure of page popularity we use the *page view* data provided by the Wikimedia Foundation.⁵ The dataset contains for each page in any Wikimedia project the number of requests per hour. We used this dataset for our study on reading preferences. For the 13-month period under consideration, we aggregated the hourly views for each month, to have monthly views for each article. The resulting dataset comprises a total of 4.3 million articles. The most visited page is the Main page, with 600 million page requests. Within this dataset we identified 1.02 million biography articles having 460 million page views in total.

Browsing data. Page popularity is only one criterion that can be considered when studying readers in Wikipedia. For example, accounting for the time spent on a page and the pages accessed during a visit on Wikipedia provides additional insights about reading activity. This information can be

⁵<http://dumps.wikimedia.org/other/pagecounts-raw/>

obtained from clickstream log data containing the entire navigation trace of users.

Since clickstream data are not provided by the Wikimedia Foundation, we collected anonymised log data for a sample of users who gave their consent to provide browsing data through the Yahoo toolbar. We identified in these browsing data users who have accessed the English Wikipedia by requesting for the following two types of URLs:

```
http://en.wikipedia.org/wiki/PAGE
http://en.wikipedia.org/w/index.php?title=PAGE
```

where *PAGE* refers to the title of the page that was viewed. We identified these page titles in Wikipedia and resolved redirects to avoid duplicate entries.⁶ We detected 288K biography articles, accessed by 387K users, and a total of 4.5M million clicks for our 13-month sample.

Article characteristics. To characterise Wikipedia articles from the editors' side, we computed their length and edit count. We retrieved these data through the Wikimedia Tool Labs.⁷ Depending on the time window of our analysis (we used several), we computed for each article its text length (the size in bytes of the last revision of the article for the given time window) and number of edits (the number of revisions of the article during that time window).

To identify articles that have been considered of high quality by the community through its internal quality assessment system, we checked for each article whether it was included in the Wikipedia lists of Featured⁸ or Good⁹ articles or assigned as an A-class article¹⁰ at the end of our 13-month period. These articles have been assessed by Wikipedia's editors using a set of pre-defined criteria developed over the course of the Wikipedia project, such as being well-written, comprehensive, and neutral. We found that 0.37% of the 1.02 million biography articles were assessed of high quality. 3% of these articles are A-class articles, 74% are good articles, and 23% are featured. In the rest of this chapter, we refer to these articles as high quality articles (*HQA*).

⁶<http://en.wikipedia.org/wiki/Wikipedia:Redirect>

⁷https://wikitech.wikimedia.org/wiki/Nova_Resource:Tools/Help

⁸http://en.wikipedia.org/wiki/Wikipedia:Featured_articles

⁹http://en.wikipedia.org/wiki/Wikipedia:Good_articles

¹⁰http://en.wikipedia.org/wiki/Wikipedia:Version_1.0_Editorial_Team/Assessment

8.4 Reading Preference

In the first part of our study we look at the *reading preferences* of users on Wikipedia. First, we identify what are the most read topics on Wikipedia and show that articles belonging to the most popular topics do not necessarily correspond to those frequently edited by Wikipedia editors. In other words, reading preferences do not always align with *editor preferences*. We then characterise the difference between reader and editor preferences using a *preference* matrix. All studies in this section are based on the page view data provided by the Wikimedia Foundation.

8.4.1 Popular Topics

In the first part of our analyses, we study the popularity of topics in Wikipedia. We select the 500 most read articles, measured by the number of article views over our data period.¹¹

We manually assigned a topic to each article using a three-round process.¹² In the first round, we collaboratively coded the articles (about 50) by using Wikipedia categories as reference point until we obtained an almost stable set of topics for these articles. In the second round, we separately coded the remaining articles. In the third round, we checked the assigned topics and discussed all ambiguous cases. To ensure a shared understanding of the existing topics, the second and third rounds were iterative. Newly introduced topics were cross-validated over the entire dataset. This process resulted into 12 distinct topics listed in the first column of Table 8.1. A description of the topics is provided in the fourth column.

From Table 8.1, we see that a large percentage of users access Wikipedia to read about entertainment-related topics such as TV series, movies, and biographies of actors and singers. Articles related to history, health and tech content (such as web services and software) are also frequently accessed. This is in accordance with previous studies [239; 260].

¹¹We selected the 500 most read articles only, as we did not observe significant changes in our results by considering more articles.

¹²The hierarchical and overlapping structure of Wikipedia's category system prevented us to automatically determine the main topic of an article in a straightforward manner.

Table 8.1: Article topics, percentages of articles in each topic, and percentage of high quality articles ($\%HQA$) in each topic for the 500 most popular articles (measured using page views).

Topics	%Articles	%HQA	Description
Biography	44.2%	31.2%	<i>Biographies of persons</i>
Media personality	18.8%	24.5%	
Musician	11.6%	37.9%	
Sportsperson	6.8%	35.3%	
Historical figure	4.2%	33.3%	
Politic./businessp.	1.8%	33.3%	
Criminal/victim	0.4%	0.0%	
Misc	0.4%	50.0%	
Publisher/writer	0.2%	100.0%	
Entertainment	17.4%	32.2%	<i>Cinema and TV</i>
Series	10.8%	22.2%	
Movie	5.4%	55.6%	
Misc	1.2%	16.7%	
List	7.6%	0.0%	<i>“List of” articles</i>
Tech	5.0%	12.0%	<i>Web, software, electronics, etc.</i>
History	4.4%	22.7%	<i>Wars, monuments, incidents, etc.</i>
Misc	3.8%	15.8%	<i>Further articles</i>
Health	3.4%	23.5%	<i>Diseases, medicine, etc.</i>
Leisure	3.2%	18.8%	<i>Games, novels, etc.</i>
Sport	3.0%	66.7%	<i>Sports, sport events, etc.</i>
Places	2.8%	21.4%	<i>Regions, buildings, etc.</i>
Adult	2.6%	7.7%	<i>Articles about adult content</i>
Culture/Belief	2.6%	7.7%	<i>Religions, festivals, etc.</i>

The third column of Table 8.1 shows the percentage of high quality articles per topic. The lower the topic popularity, the smaller the number of high quality articles belonging to that topic. Indeed, we observe a Spearman’s rho coefficient of $\rho = 0.72$ (p -value $\ll 0.01$), suggesting a high correlation between topic popularity and the percentage of high quality articles. However, there are some exceptions. For instance, for the topics “Health” and “Sport”, although the percentage of articles belonging to these topics is relatively low, many articles are of good or high quality (23.5%, and 66.7%, respectively). On the other hand, the percentage of high quality articles in the “Tech” area is low (12.0%), albeit this being the fourth most popular topic in our dataset.

These observations suggest some degree of non-alignment between *users’* reading preferences and *authors’* editing preferences. To examine this further, we define several measures to characterise these two preferences next.

8.4.2 Reading and Editing Preferences

Table 8.1 provides a first indication of some non-alignment between reading and editing preferences in Wikipedia. In this section, we define various measures to study this.

Measuring reading preferences. We determine readers' preferences using an adaption of the popularity metric *#Clicks* of Chapter 4. Whereas *#Clicks* measures the number of page views on a site, we measure here the popularity of articles by the number of page views to that article. Previous studies suggest that popularity is a dynamic phenomenon that can partly be characterised by bursty behaviour of page views [214; 248]. Our goal is to determine a value that best represents the popularity of an article by filtering out such bursty behaviour. Thus, we calculate the monthly article popularity measured by the number of page views in each month from September 2011 to September 2012. Then, we measure the median rank of article popularity ($Popularity_a$) by their monthly popularity, which is less sensitive to outliers.

Measuring editing preferences. To determine editors' preferences, *i.e.* the articles they are mostly working on, we use three measures, each indicating a particular angle regarding editors' preferences. First, we employ the number of edits ($\#Edits_a$), a common measure of editing activity. For each article, we calculate the number of revisions over the whole period range. This measure, however, does not provide information about the effect of an edit, such as its informativeness and quality. We therefore propose to use article length ($ArticleLength_a$) as a measure for the informativeness of an article. The fact that an article is long suggests that a number of editors spent time and effort writing about the topic of the article, to make it more informative. We calculate the length of an article for a given time period using the latest version of the article in that period. Finally, editing may lead to the article being identified by the community as good, featured, or A-class (the pinnacle of the editing process). This would happen when the article is considered to provide comprehensive information on a topic.¹³ We use the available data provided by Wikipedia – whether an article is a good or featured article, or belongs to the A-class articles (HQA_a) at the end of our data period – as a measure for article quality.

¹³See criteria for featured articles http://en.wikipedia.org/wiki/Wikipedia:Featured_article_criteria

We compare reader and editor preferences by measuring the correlation between the reader preference measure $Popularity_a$ and the editor preference measures $ArticleLength_a$ and $\#Edits_a$. As discussed in Section 8.3, we focus on biography articles, which form the most popular article topic in Table 8.1. We report Spearman’s rho coefficient (ρ) for the metrics. We observe low correlations: 0.22 for $ArticleLength_a$, and 0.16 for $\#Edits_a$. These values suggest some non-alignment between reader and editor preferences. To further investigate this, we built a linear regression model using $ArticleLength_a$, $\#Edits_a$, and HQA_a as features to predict the number of page views of an article. Our model predicted the number of page views with a coefficient of determination of $R^2 = 0.24$ ($R^2 = 1.0$ would represent a perfect fitting model), further indicating that readers and editors preferences diverge in many cases.

Next, we introduce a *preference* matrix, which allows visualizing the differences in reading and editing preferences using the above defined measures.

8.4.3 Preference Matrix

For each article, we calculate its popularity (our reading preference measure) and its length (an editing preference measure).

The distributions of popularity values and article length values indicate whether articles are popular or not, and whether articles are long or short. We determine the upper and lower quartiles of both distributions since we want to identify articles with extreme values. We remove all articles that fall into the interquartile range of the article length or popularity distribution (the middle 25 – 75% of both distributions). This means that we only consider articles that differ significantly from those having an average length or popularity.

This results in the four groups of articles shown in Figure 8.1. The horizontal axis represents article popularity (the reading preference) and the vertical axis represents article length (one of the measures characterizing editing preference). The values of both measures are transformed into an ordinal scale to overcome scaling issues, *i.e.* we ranked all values for article popularity and article length. Each dot in the matrix represents an article and the position corresponds to its popularity and length. We only show a random sample of 100 articles in Figure 8.1 to improve legibility.

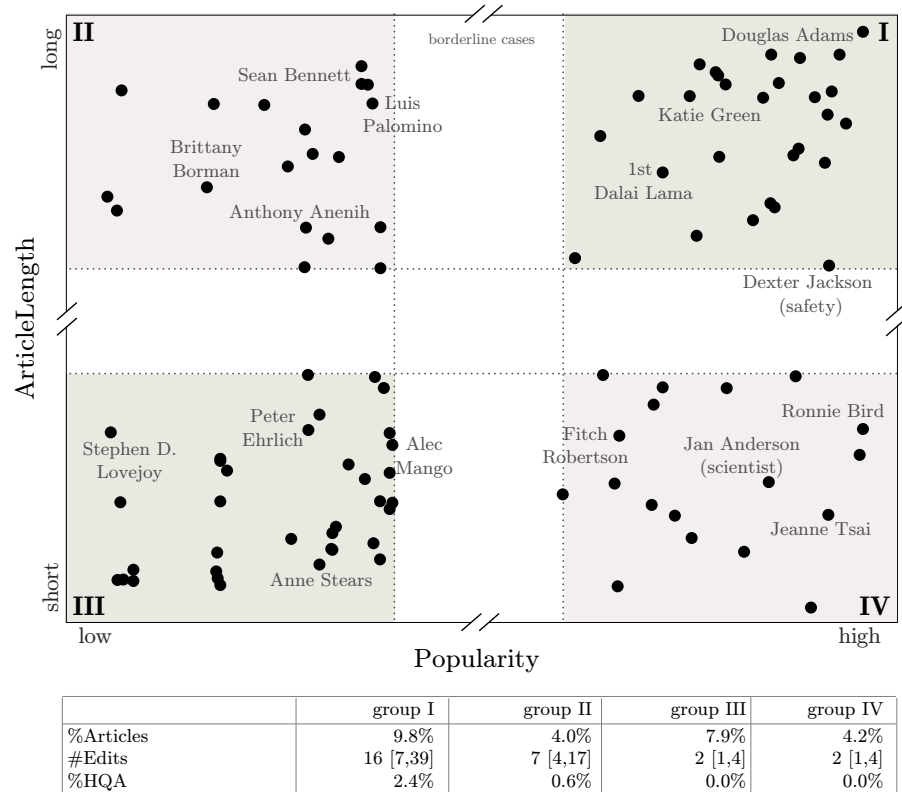


Figure 8.1: Preference matrix: (Top) Preference matrix defined by article popularity and length. (Bottom) Percentage of articles belonging to a group (*%Articles*), median and interquartile range of the number of edits per group over the whole data range (*#Edits*), and percentage of high quality articles in each group (*%HQA*).

Under the preference matrix, we report the percentage of articles belonging to each group, and the other two editing preference measures, namely, the percentage of high quality articles, and the median and interquartile range of the number of edits. We see that featured articles tend to be long, confirming previous work [273] and suggesting a relationship between article length and article quality.

Many articles belong to group *I* (9.8%) and group *III* (7.9%). Whereas group *I* contains very long and often read articles, articles in group *III* are short and seldom read. In both groups, we have articles for which editing and reading preferences align.

A divergence between reader and editor preferences can be observed for articles belonging to groups *II* and *IV*. Group *II* articles (4% of all articles) tend to be not read very often, even though they are very long (probably very informative). This group also contains a low number of high quality articles. For instance, it contains the biographies of the Nigerian politician “Anthony Anenih” and the American football player “Sean Bennett”. We speculate that not many users read these articles because the person in question is not popular nowadays (*e.g.* former American football player) or is of interest only to a specific user community (*e.g.* users interested in Nigerian politics). This is further accentuated by the lower edit activity in this group (median of 7 edits) compared to group *I* (median of 16 edits). In fact, many of the articles forming group *II* are on topics that were popular in the past and heavily edited during that time.

Finally, 4.2% of all articles belong to group *IV*. For these articles as well, reader and editor preferences do not align. Even though articles are regularly accessed by readers, they are short (and have seldom been edited) and none of them is of high quality. Taking the examples of “Jan Anderson (scientist)” and “Ronnie Bird”, we see that these articles are often viewed, but are short and have hardly been edited during the last 13 months (median of 3 edits per article). Additionally, none of these articles is considered to be of high quality, even though readers access them very often.

To summarise, we observe differences between what readers access and what editors work on. The most edited articles tend to be long (groups *I* and *II*) and the number of high quality articles in these groups is higher compared to the other two groups. However, only articles in group *I* are very popular, suggesting that article quality does not drive popularity.

The opposite can be observed for articles in groups *III* and *IV*. These groups contain shorter articles, and fewer high quality articles. Moreover, articles in these groups tend to be edited less. This indicates that editors rarely added content to them in the past, reflecting low interests in these articles. Whereas articles in group *III* neither meet authors nor readers interests, we can see that readers are interested in articles of group *IV* despite the scarce attention they receive from editors.

Next, we analyze how users read articles during their online sessions, and how this matches with the editing activity.

8.5 Reading Behaviour

Here we study how users read Wikipedia articles. That is, we look at the biographies that users read when *visiting* Wikipedia. We introduce three measures to characterise patterns of reading behaviour: $ArticleViews_a$, $ReadingTime_a$ and $SessionArticles_a$.

Measuring reading behaviour. All reading activities of a user on Wikipedia during an online session form what we refer to as a *reading session*.¹⁴ During a reading session, a user spends time reading an article a . We use $ReadingTime_a$ to refer to the time spent on article a . The metric is an adaption of the site engagement metric $DwellTimeS$ of Chapter 5 on Wikipedia articles. This user may return to article a several times during a session, and visit several other articles. Following Chapter 5, we refer to this type of behaviour as *online multitasking*, and in this chapter we use the multitasking metrics $SessVisits$ and $SessSites$ to study how users access and re-access articles in Wikipedia. We employ the metrics as follows: $ArticleViews_a$ is the number of times the article a was viewed (known as $SessVisits$) and $SessionArticles_a$ is the number of articles viewed during the reading session where article a was read (known as $SessSites$).

Data processing. We use the browsing data as it enables us to access the readers' entire navigation traces (Wikipedia articles and other web pages) during their online sessions. To have a more homogeneous and robust dataset, we discard articles with lower values of length or popularity, and focused our analysis on articles belonging to group I of the preference matrix (see Figure 8.1), which contains the large majority of articles in our browsing data (83.47%). These articles allow for a reliable interpretation of any observed difference between reading interests and editing preferences since their length and popularity are high enough.

We characterise the reading behaviour of an article a by calculating per month the average of $ArticleViews_a$, $ReadingTime_a$ and $SessionArticles_a$. We also calculate $Popularity_a$, the popularity measure defined in the previous section. Therefore, for each article a we obtain 13 vectors, one for each month of the 13-month period. We refer to each vector ($ArticleViews_a$, $ReadingTime_a$, $SessionArticles_a$, $Popularity_a$) as a *behaviour vector*.

We generate behaviour vectors of an article for the months where it was visited in at least 10 reading sessions. This enables us to derive stable

¹⁴See Section 3.2 for a detailed definition of online sessions.

values for the three measures calculated based on reading sessions. This results into 9,726 articles and 49,921 behaviour vectors. To ensure that the two datasets (page view and browsing data) are comparable, *i.e.* no strong bias in the browsing data is influencing our results, we ranked the articles according to their overall popularity in both datasets, and found that their rankings correlate (Spearman's $\rho = 0.64$).

8.5.1 Reading Patterns

To detect patterns of reading behaviour we use the approach introduced in Section 4.5. We use the k-means algorithm to cluster the behaviour vectors, whereas the values of each measure are transformed into an ordinal scale to overcome scaling issues. The number of clusters is determined by a minimal cluster size such that each cluster contains at least 20% of the 49,921 behaviour vectors. Since the clustering is performed with the behaviour vectors of the articles, an article can occur in multiple clusters. This allows us to analyze changes in the reading pattern of an article across the 13-month period; we return to this in Section 8.5.2.

We obtain four clusters, shown in Figure 8.2, each corresponding to a pattern of reading behaviour. The first row displays the name given to each cluster. The second row contains the cluster centers normalised by the z-score, hence the plots show the extent to which the standard deviation of a metric rank is above or below the mean. The third row contains the number of articles and behaviour vectors within each cluster. Since the sizes of the clusters are similar, there is no dominant reading pattern. The fourth row shows the predominant article subtopics per cluster measured by the probability difference PD .¹⁵ The measure describes the likelihood that a subtopic occurs in a given cluster with respect to its likelihood that it occurs at all (see Section 3.3 for a detailed definition of PD). We only show the subtopics with the largest PD values. The last three rows of Figure 8.2 report the values of the three editing preference measures. For each behaviour vector, we calculated the length of the corresponding article, using the latest revision of the article for the given month, and the number of edits made during the month (we report median and interquartile range). We also determined the percentage of high quality articles.

¹⁵For each cluster, we sampled at random a subset of 500 articles, and determined the sub-categories of these articles by using the three-round process described in Section 8.4.1. We manually categorised all articles based on the subtopics of the category biography as shown in Table 8.1.

We discuss now each of the identified reading patterns and relate them to the editing preferences. The patterns “Focus”, “Trending”, and “Exploration” are what content portals aspire to: users spending time reading their articles and/or reading many articles.

Focus. Articles following this pattern are characterised by an expected encyclopedic reading behaviour: people spend a lot of time reading the article (high $ReadingTime_a$), but access very few other articles (low value of $SessionArticles_a$) within the session. Users have a specific information need (*e.g.* they want to learn something about “Jacques Cousteau”). Articles in this cluster have a lower than average popularity, and are more likely about artists/writers, historical figures, and politicians/businesspersons.

The high reading time indicates a strong interest in the content of the article. Hence, we would expect many of these articles to be marked as good, featured, or A-class, because the quality of these articles seems important. However, the percentage of high quality articles in “Focus” ($\%HQA = 7.7\%$) is lower than for the “Trending” and “Exploration” clusters. Moreover, although we observe an appropriate article length ($ArtLen = 28K$), the number of edits ($\#Edits = 11$) suggests that editors are not interested in improving these articles. Indeed, the article about “Jacques Cousteau” is long (a median of $30K$ characters), but it is neither featured nor good nor A-class, and the number of edits is low (a median of 5.5 edits per month).

Trending. Many biographies about historical figures, musicians and criminals/victims follow this pattern: articles are visited very often (high $Popularity_a$). Users read only a few other articles (low $SessionArticles_a$), similarly to the “Focus” reading pattern, but they spend less time reading the articles. This suggests that users are probably “quickly looking up” for information about something that is currently trending or has recently happened. For example, users read about the politician “Ron Paul” when he was a candidate for the presidency of the United States, but only to catch up on any recent news about him.

“Trending” articles exhibit the highest edit activity and the highest percentage of high quality articles compared to the other two clusters ($\#Edits = 20$ and $\%HQA = 16.9\%$). These articles not only attract users to read them but also authors to edit them, which is in accordance with a previous study by Reinoso [216], and also aligns with the work from Keegan *et al.* [123] about breaking news and current events in Wikipedia. The high percentage of high quality articles suggests that editors do not only work on the articles to increase the quality, but also to “update” information caused by recent or

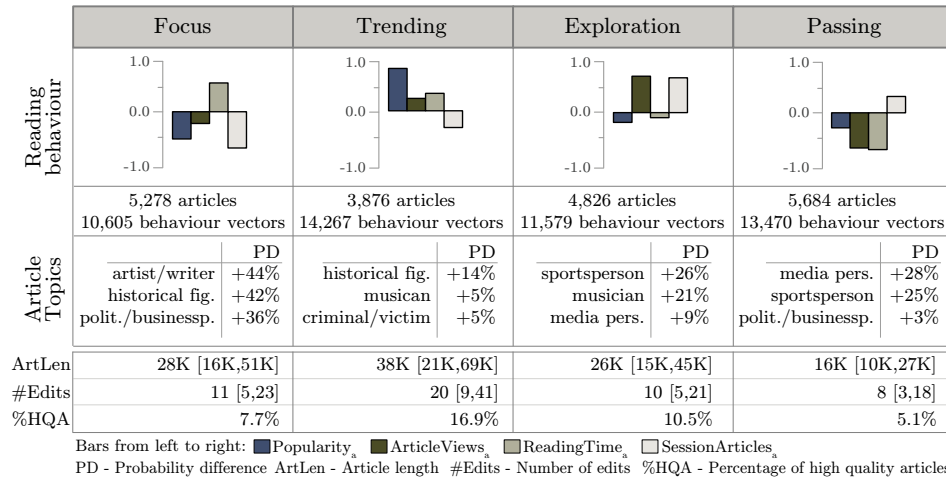


Figure 8.2: Reading patterns: (1st row) Article clusters and reading characteristics. (2nd row) Number of articles and behaviour vectors per cluster. (3rd row) Predominant article topics per cluster. (4th row) Median and interquartile range of the article length (*ArtLen*), number of edits per cluster (*#Edits*), and percentage of high quality articles in each group (*%HQA*).

continuous events related to the article topic. Indeed, we saw in our dataset that featured articles are also edited frequently (a median of 19 edits per month). Featured articles are usually only changed in case new information becomes available.

Returning to our previous example, the politician “Ron Paul”, we observed a median of 81 edits per months during the time the article was trending (December 2011 until May 2012). In the other months (when “Ron Paul” was not competing for the presidential primaries), the article belonged to the “Focus” cluster and had only 20 edits per month. We return to this later in this section.

Exploration. This pattern predominantly contains biographies describing sportspersons, musicians, and media personalities that have an average popularity. The number of articles viewed in a session (*SessionArticles_a*) is the highest compared to the other clusters, indicating that users explore many other articles in a reading session. Looking into the articles that were visited, we saw that articles requested during the same session belong mostly to the same topic (*e.g.* users who read the article about the actor “Al Pacino”

also read articles about his movies).¹⁶ The high value of $ArticleViews_a$ indicates that users return regularly to the article under consideration, suggesting that they use it as a basis to navigate to other articles on the same topic. This hypothesis is supported by the low reading time of the focal article.

The editing preferences are comparable to the “Focus” pattern, in terms of number of edits (moderate values of 10 edits per article) and article length. The difference between the “Focus” and the “Exploration” reading patterns may be explained by external factors that influence the consumption of online content by users, such as the death of a famous artist [214].

Passing. Many biography articles about media personalities, sportspersons, and politicians/businesspersons belong to this cluster. The number of articles viewed in a session ($SessionArticles_a$) is above average, suggesting that users read different articles. Users browse many articles in the same session, but in contrast to “Exploration” they seem to only pass through the focal article (low $ReadingTime_a$), and do not return to it (low $ArticleViews_a$).

An example is the article about “Jackie Jackson”, member of “The Jackson 5”. When users are reading about “The Jacksons”, they also view this article, but then quickly move to other related articles. The question is whether users do not spend much time on the article, because they are not interested in reading more about “Jackie Jackson”, or because there is not much information provided about her (her article has a median text length of 9K).

Indeed, compared to the other clusters, the “Passing” cluster has a lower percentage of high quality articles ($\%HQA = 5.1\%$), and has shorter articles ($ArtLen = 16K$) and the lowest number of edits per article (a median of 8).

To summarise, we observe that articles exhibit different reading patterns. These seem to be mainly driven by the topics of the articles and therefore the interests of users, and less by their quality. Users show their interest in an article in different ways, for instance, by exploring also related articles (“Exploration” cluster) or by spending time reading the article (“Focus” cluster). Sometimes, the interest in an article is driven by external factors, as shown with articles belonging to the “Trending” cluster (*e.g.* users read

¹⁶We extracted all wikilinks between the articles in each reading session and found that on average over 76% of the articles visited in a session are connected to one another. This applies even for long reading sessions containing more than 10 articles (the average becomes 70%).

biographies about currently trending persons). On the other hand, for articles belonging to the “Passing” cluster, the question is whether the reading behaviour is partly caused by a lower quality of the articles. Increasing the article quality might lead to readers getting more engagement with the article. Overall, our results show that popularity and reading time are not the only factors that should be taken into account when studying user reading behaviour – multitasking metrics provide further information about how users read articles in Wikipedia; more precise, how often users return to an article and how many other related articles they read.

Three out of the four clusters constitute reading patterns where users are interested in the articles they are reading. However, editors seem to focus on articles in mainly one cluster, the “Trending cluster”. The editing activity, the article length and the percentage of high quality articles is higher in that cluster than in the “Focus” and “Exploration” cluster. This shows again a non-alignment between reader and editor preferences.

Finally, as shown in our example of “Ron Paul”, an article can be in several clusters, depending on the month under consideration. That is, articles can transition between patterns across the 13-month period. We study this next.

8.5.2 Changes in Reading Patterns

The analysis conducted in the previous section used measures calculated on a monthly basis (the behaviour vectors) to identify reading patterns. As a result, articles can belong to more than one cluster. In this section, we use this fact to study how articles might move (if they do) between reading patterns, and discuss possible reasons for these transitions. First, we determine how stable articles are in terms of their popularity and the way they are read across the 13-month period. We then look at typical transitions between reading patterns.

Stability

We calculate the number of months in which an article was visited in at least 10 reading sessions. We refer to this as the article *longevity*, denoted $Longevity_a$ for article a . In Figure 8.3a, we plot on the x-axis the longevity values and on the y-axis the percentage of articles for a given longevity value. Almost 30% of the articles (2,836) have a longevity value of 1, meaning that these articles have been accessed in at least 10 reading sessions only in a single one-month period. Another 13% of the articles (1,264) have been

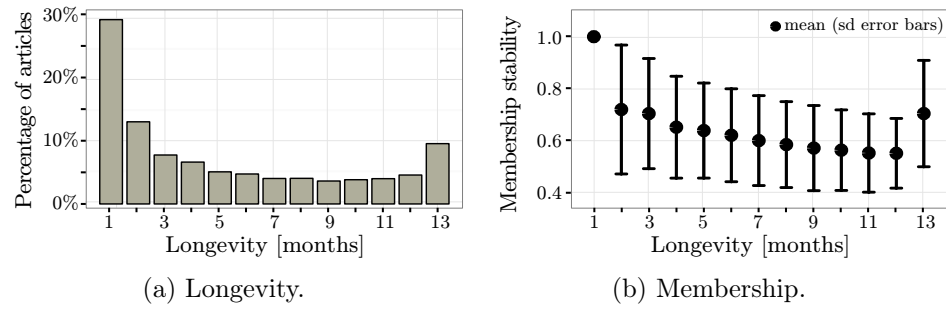


Figure 8.3: Stability of articles: For each longevity value, we plot the percentage of articles with that value, and the mean and standard deviation of the membership values of the articles.

accessed in at least 10 reading sessions in two different months. This percentage decreases continuously for larger numbers of months, but increases again for 11 and more months. About 10% (928) of articles are read at least 10 times a month over the whole 13-month period. This suggests that there are articles that are frequently accessed over a long time period.

We examine the stability of an article in terms of which clusters it belongs to (*i.e.* the reading patterns it exhibits). We calculate the number of months an article a remains in its “home” cluster, which is the predominant reading pattern exhibited by the article. Then, we normalise this value by dividing it with the corresponding $Longevity_a$ value. We refer to this as the article *membership* stability, denoted $Membership_a$ for article a . In Figure 8.3b, the y-axis shows the average and standard deviation of the *membership* values for all articles for a given longevity. For example, an article with a longevity of 3 (the article was visited in at least 10 reading sessions during three, not necessarily consecutive, months) has a membership value of 0.7 on average. This means that on average the article was read 70% of its lifetime according to its most frequent reading pattern.

Figure 8.3b suggests that the higher an article’s longevity, the lower its membership stability. This means that the longer - in terms of months - the article is accessed frequently, the higher the probability that its reading pattern changes. However, the average membership stability values are always above 0.5, indicating that many articles remain in their “home” cluster for at least 50% of their lifetime. It is interesting to note that the membership stability increases again for articles with a longevity value of 13. This means that high longevity implies high membership stability.

Transitions

We study the most frequent changes, *i.e.* transitions, between reading patterns (clusters) and explore possible reasons behind these changes. A $Transition_a$ exists for an article a if one behaviour vector of a belongs to cluster C at month m and another behaviour vector belongs to cluster D at month $m + 1$, where the clusters represent two distinct reading patterns. We selected two cases to explore transitions. We consider all articles and then only articles with a $Longevity_a$ value of 13 – the set of highly stable articles in terms of their monthly access rate.

In Figure 8.4, we visualise the transitions between the four clusters by two networks, one for each case. Each vertex represents one cluster (*i.e.* reading pattern) and the size of a vertex corresponds to the number of articles in that cluster. The undirected edges in the network depict the transitions between the clusters. We use an undirected network since we observed a similar number of transitions in both directions. The largest difference we observed is smaller than 2.0%, which can be explained by the fact that an article usually belongs to one cluster (*e.g.* “Exploration”), moves to another cluster for a short time (*e.g.* to “Trending” because something happened with the person under consideration), and then moves back to the original cluster.

Each edge has a weight, which is the percentual amount of transitions between two clusters; for example, an edge weight of 23% means that 23% of all transitions in the network take place between these two clusters.

The complete network (left side of Figure 8.4) show how external factors, such as recent or continuous events related to a person, drive changes in reading patterns. This is the case for example for the biography article of the Facebook co-founder “Chris Hughes”. Before March 2012, users tended to “pass by” this article (when reading about Facebook). However, this changed in March 2012 when Chris Hughes became the owner of the “The New Republic” magazine, attracting some media attention. Users started reading this article in a more “explorative” manner, using it as a starting point to access other articles related to the person.

The edge weights differ a lot in the network. We see a strong connection between the “Passing”, “Exploration”, and “Trending” clusters, indicating that many articles adopt all three reading patterns and sway between clusters. A transition can be even long-lasting, as in the case of the article “Jacqueline Kennedy Onassis”. Until April 2012, the article was in the “Trending”

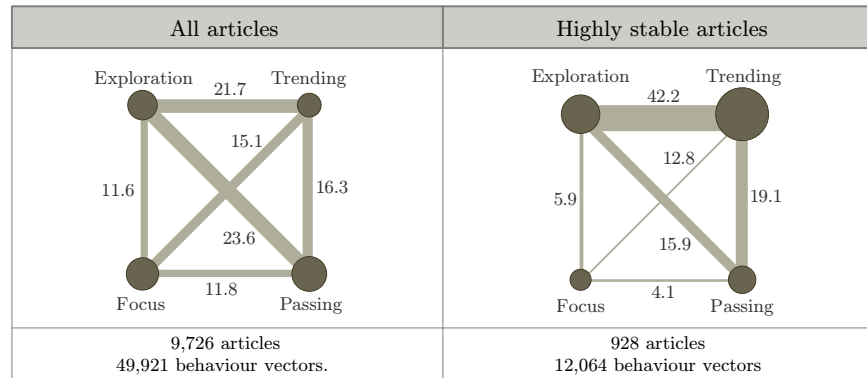


Figure 8.4: Transitions between reading patterns considering all articles (left) and only articles with a high stability (right). The vertex size represents the number of articles that belong to the cluster, and the edge weight represents the percentage of transitions between two clusters.

cluster, but then lost its popularity and moved to the “Exploration” cluster. We assume that the article started to trend when her audio tapes, recorded after her husband’s assassination, were released.

We also observe that articles belonging to the “Focus” cluster are isolated – the likelihood that an article is moving from or to the “Focus” cluster is low. Articles that are read in this way can be considered as the most stable ones, as their reading behaviour hardly varies. An example of such an article is “Franklin D. Roosevelt”.

Looking at stable articles only (right side of Figure 8.4) we see a different pattern. Compared to the network comprising all articles, we observe that “Focus” becomes even more isolated, showing again its special characteristic – a constant reading pattern. The transitions between “Exploration” and “Trending” become stronger, whereas the transitions between “Exploration” and “Passing” become weaker. This implies that the former two clusters indeed describe reading patterns for the same type of articles.

8.6 Discussion

We analyzed readers’ preferences and reading behaviour. We did so by connecting them to editors’ preferences, allowing us to relate the usage side of Wikipedia to its production side. Our goal was to provide insights about how the reading experience and the editing process on Wikipedia could be

enhanced. We discuss now our main results, position them in light of our goal, and present examples of potential applications.

Reading preferences. Using the page view data provided by the Wikimedia Foundation, we studied reading preferences of users on Wikipedia. Our results confirm other works showing the dominance of *entertainment-related topics* among the most read topics on Wikipedia [239; 260]. The encyclopedic character of Wikipedia does not exempt it from following the known prominence of consuming and interacting with entertainment-related content observed on the Web.

We then introduced a preference matrix, which enabled us to differentiate four groups of reading versus editing preferences. These groups provide valuable insights to Wikipedia’s quality system, in particular groups *II* and *IV*, where the preferences do not align.

Group *II* articles are often edited but not often read, whereas group *IV* contains articles that are popular, but hardly looked at by editors. Being aware of these divergences can help Wikipedia editors making an *informed* decision about which articles to focus next. As opposed to tools such as WikiDashboard [244], which allows readers to evaluate article quality on the basis of an author history, the preference matrix can provide editors with a visualisation of user reading preferences. This might draw their attention to articles or topics they have not edited before. Moreover, task recommendation services, such as the SuggestBot [51], could use the preference matrix as input to recommending tasks. In addition, improving articles that are of readers’ interest might engage them more with Wikipedia.

Reading behaviour. In the second part of our work, we studied the reading behaviour of users by adapting site engagement and multitasking metrics to our context. We could show that, in addition to article popularity and reading time, multitasking metrics further enhance our understanding on how users read articles in Wikipedia. Multitasking metrics, although focused on one article, take into account the reading behaviour within the whole reading session by addressing how often users return to the focal article and how many other related articles are read. Using the pattern detection approach of Section 4.5, we clustered the articles characterised by these metrics and we identified four main reading patterns: “Focus”, “Trending”, “Exploration” and “Passing”.

Information about the reading behaviour of users can be useful in many ways, such as for the selection of articles for the main page, or the Article

Feedback Tool (AFT).¹⁷ Knowing which articles follow, for instance, the “Focus” pattern might help making the Article Feedback Tool more efficient since using this tool over the entire Wikipedia corpus failed. Editors complained about the low quality of the feedback made on articles. The fact that an article is often read and users spent time on it may indicate that users are interested in the article. As such, their feedback (if any) is likely to be more constructive and valuable.

In conformity with the work from Gyllstrom *et al.* [94], we showed that the reading behaviour depends less on the article quality, but more on the article topic and therefore the interests of the reader. The quality of articles does not greatly influence what users choose to read. In general, the editing activity and the quality of articles reflect mostly the authors’ interests and not the readers’ interests. The exception to this are articles belonging to the “Trending” pattern, which are both accessed by many users and edited by many authors, compared to the three other reading patterns.

Understanding reader preferences and behaviour can support editors in their work in several ways. The identified browsing patterns provide information about which articles the readers are interested in and how these and related content are read. If readers are interested in an article topic, they tend to look up information (“Trending”), spend a lot of time in reading the article (“Focus”), or consume the article content, but also related information (“Exploration”).

This information can be used to improve the structure and presentation of the article content. For instance, the “Exploration” pattern corresponds to a navigation “way” to consume Wikipedia content. One article is focal, but also acts as a source to explore other articles. Knowing that these articles are consumed in this way, Wikipedia editors may add more links, keeping the users engaged by providing additional and relevant content. From an interface design perspective, navigation tools could be provided to guide users with the aim to enhance their reading experience.

Additionally, reading pattern can help editors to decide which articles to edit next. For instance, for the “Focus” pattern, we observed the highest reading time per article compared to all other patterns, but the percentage of high quality articles is lower than in the “Trending” and “Exploration” cluster. With respect to Wikipedia’s production side, articles following the “Focus” pattern may greatly benefit from improvements in their quality, as users are very interested in them.

¹⁷http://en.wikipedia.org/wiki/Wikipedia:Article_Feedback_Tool

Also articles belonging to the “Passing” cluster may benefit from improvements. We assumed that users are not interested in the article, and therefore only pass through the focal article during their reading session. Another explanation is that these articles are not very informative (they are often short), and have rarely (likely as a consequence) been marked as good or chosen to be featured.

Stability. Finally, we looked at the stability of the reading patterns. We found that many articles are stable (remain in the same cluster), and that changes of the reading patterns are of temporal nature (*e.g.* in case of an event) or due to the time passing (*e.g.* interest in the person is decreasing). Studying the transitions between the reading patterns revealed two main findings. First, we observed a strong connection between the “Exploration” and “Trending” clusters, indicating that many articles adopt both reading patterns. Second, we observed that the “Focus” cluster represents a reading pattern that is isolated from the others. Articles in this cluster usually do not change their reading pattern. It indicates that this pattern represents articles with a high stability.

The above observations can inform the Wikipedia editor community in two ways. The stability of articles allows them to make long-lasting decisions for their editorial work. For instance, when adapting an article for explorative reading, these adaptations, such as adding links, are useful for the consumption of that article later on. On the other hand, transitions between reading patterns inform editors about recent trends (*e.g.* when an article is moving from “Passing” to “Exploration”, indicating an increased interest from the reader side). Such articles can be candidates to be placed on the front page to raise awareness.

Concluding remarks. This research provides new insights about how users consume content on Wikipedia: their reading preferences and behaviour. This research also attempts to connect Wikipedia’s readers (usage side) and Wikipedia’s editors (production side). Using several measures to characterise reading preferences and behaviour, we learn how users consume Wikipedia content, and illustrate how this information could inform Wikipedia editors about their editing tasks, for instance which articles to prioritise and why.

Identifying how an article is read can be used to determine which articles are more “engaging” than others or which articles need to be improved with respect to their “engagement”, for instance, as measured by the average time

spent on the article or the number of articles accessed from it. Articles that are more engaging are likely to promote a successful reading experience and even encourage users to return to them or to other articles. Readers that regularly return to Wikipedia are more likely to recognise the effort of Wikipedia's community and might even develop a sense of belonging to that community [210]. This in itself may further engage Wikipedia editors since they feel that their work is recognised and appreciated.

Limitations. Our work has a number of limitations. First, the results of this work only present readers' preferences and behaviours on biography articles. It is, however, maybe the case that articles related to, for instance, places¹⁸ or mathematics¹⁹, or "list of"²⁰ articles exhibit a very different structure. If we would have considered all articles in our study, the reading patterns would only reflect the structural differences between different types of articles. We therefore decided to focus on biography articles; the most popular topic in Wikipedia. However, we might obtain different results when using articles related to other topics. In order to facilitate this, a method that automatically determines the topics of articles should be developed, because our approach is limited by the fact that our topics are manually defined.

Second, the fact that we only consider a subset of the whole browsing data on Wikipedia limits our results related to the longevity of articles. Since some articles are only popular over a certain period of time, the reading activity on these articles is not available in our dataset when they become unpopular. This information might be available for some articles in the whole browsing dataset of Wikipedia. However, these data are not available, and therefore it was necessary to focus on a subset of it in our study.

Finally, we only employed a small set of features that characterise the articles from an editor perspective. Information about the number of links and headings in an article, or the completeness and complexity of the article text would provide further insights about the article characteristics and how it might affect readers' engagement.

¹⁸<http://en.wikipedia.org/wiki/Peru>

¹⁹http://en.wikipedia.org/wiki/Discounted_cumulative_gain

²⁰http://en.wikipedia.org/wiki/List_of_scientific_journals



PART IV

Applications II: Inter-Site Engagement

This part contains two case studies on inter-site engagement, which are concerned with engagement in a provider network and a network of news sites. For the latter study, we develop in the last chapter an application that has the potential to increase the engagement with a news site.



Engagement in a Provider Network

This chapter aims to extend our knowledge of inter-site engagement within a provider network by exploring several aspects of this type of engagement.

9.1 Introduction

User engagement has been the focus of many studies. These studies aim to understand how users engage with a website [16; 193], and which factors influence it [67; 147], leading to a deeper understanding of how we can increase engagement [84]. However, there is limited knowledge about how users engage with a provider network that offers a variety of sites encompassing diverse services such as news, mail, and search. In Chapter 6, we have tackled this problem by defining various metrics that enable us to measure inter-site engagement in a provider network. In this chapter, we extend this study in various ways.

We have observed that site engagement depends on the loyalty of users and whether the site is visited during the week or on the weekend (see Chapter 4). In this chapter, we study whether and how these factors influence inter-site engagement and which implications this has. In addition, we investigate how returning traffic (users leaving the network but returning within the same session), and the upstream traffic affect inter-site engagement. It has been shown [251] that user browsing activity on a site depends on the upstream traffic type, that is, where the user is coming from when entering the site. In this work, we investigate whether the same applies for inter-site engagement.

Table 9.1: Network instances based on 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.

Type	Network	Number of instances	Clicks per network
Month	US_{Feb}	1	235M
Daily	$US_{01/08}, \dots, US_{31/07}$	356	8.6M 2.9M

In our first study on inter-site engagement, we have shown that there is a strong relationship between site popularity, and the incoming and outgoing traffic of the site (see Section 6.4). We refer to this as the network effect. However, so far we do not know the extent of the network effect, that is, whether also the engagement of sites in a provider network depend on each other. We therefore study in this chapter how the traffic is distributed over the network, and how this affects the engagement of the sites.

Existing studies have shown that hyperlinks are a powerful tool to direct users through the Web [254; 279]. This implies that hyperlinks are probably very important in the context of inter-site engagement. We investigate this by analysing how the hyperlink structure of the sites in a provider network influence the inter-site and site engagement.

A review of existing research is included in Chapter 6. We describe our dataset and networks in Section 9.2. Section 9.3 analyses how the loyalty of users, day of week, and other aspects affect inter-site engagement. The network effect is investigated in Section 9.4. Section 9.5 analyses how hyperlinks influence inter-site engagement. The last section discusses our findings.

9.2 Dataset and Networks

Based on a sample of users who gave their consent to provide browsing data through the toolbar of Yahoo, we collected 12 months (August 2013 to July 2014) of anonymised interaction data. The data consists of 610M sessions. We created various provider network instances using the browsing activity of users on 73 Yahoo sites based in the Unites States. A detailed definition of provider networks is given in Section 6.3. The network instances are listed in Table 9.1.

The first provider network instance consists of the browsing data of February 2014 (US_{Feb}). The other network instances represent the browsing activity

for each day between August 2013 and July 2014, that is, we defined 365 networks ($US_{01/08}, \dots, US_{31/07}$), each representing the traffic of the network on a certain day (08 refers to August, 07 to July, *etc.*).

Site categories. We use the categorisation schema as employed in the fundamental study on inter-site engagement (see Section 6.3):

- 11% front pages and site maps [Front page]
- 23% service sites [Service]
- 25% news sites [News]
- 16% leisure and social media sites [Leisure]
- 25% provider sites [Provider]

9.3 Diversity in Inter-site Engagement

In Chapter 4, we could see that site engagement depends on the loyalty of users, and whether the user visits the network on a weekday or the weekend. In this section we study whether the same can be observed with respect to inter-site engagement. In addition, we investigate the effect of the upstream traffic type, and 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 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 9.1). 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.

9.3.1 User Loyalty

Following Chapter 4, we group users according to the number of days they have visited the network within February 2014. We simplify the loyalty levels of Chapter 4 as follows:

- Casual: 1 active day
- Active: 2-14 active days
- Passionate: more than 14 active days

We then use the browsing data of February 2014 of the Casual, Active, and Passionate users to create 3 user-based networks.

We first analyse the differences at network level (Figure 9.1 (top part)). We see that Active users navigate more often (*Flow*: +17.8%), and between more sites (*Density*: +43.1%) than Casual users. The values increase again from Active to Passionate users. This shows that the inter-site engagement increases, the more loyal the users are. Although we reported in Section 6.6 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 Active to Passionate: -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, Active users always return to the front pages to access other sites in the network (*e.g.* *frontpage* → *news* → *frontpage* → *leisure*). Passionate users, on the other hand, are “aware” of links that allow them to access other sites directly (*e.g.* *frontpage* → *news* → *leisure*). We also observe that the engagement in the network increases with the loyalty of users. Passionate 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 9.1 (bottom part) compares the Active and Passionate networks. As already observed at network-level, Passionate 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 Passionate users are rarely visiting provider-related sites. They do not need to access help sites

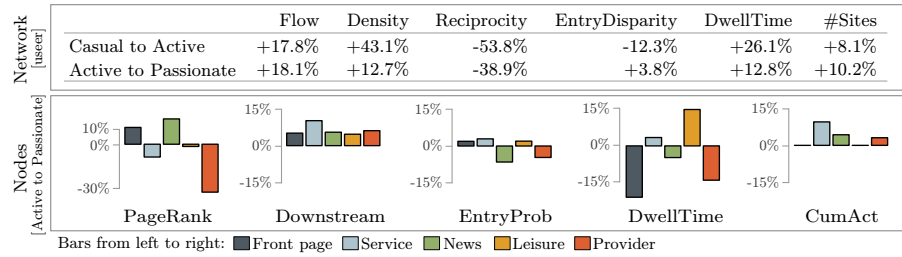


Figure 9.1: Differences between user-based networks: (Top) Network-level metrics. (Bottom) Node-level metrics.

very much, likely because their account-related settings have already been performed a while ago.

Finally, in terms of time spent on the network, Passionate users seem to spend a considerable amount of time on leisure sites (increase in *DwellTime*), compared to Active users. We also observe a significant decrease of dwell time for front pages. Front pages can be compared to search sites (see Section 4.5.1); a low dwell time is a sign of a good user experience, since these sites are used to navigate (quickly) to other sites. We speculate that Passionate users know the front page well, and hence are able to move quickly to the site they want to reach. Active 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.

9.3.2 Weekdays versus Weekend

In Chapter 4 we showed that it is important to consider temporal aspects when studying user engagement. We therefore compare inter-site engagement during weekdays with that from weekends. We use the daily networks ($US_{01/08}, \dots, US_{31/07}$) 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 9.2). 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

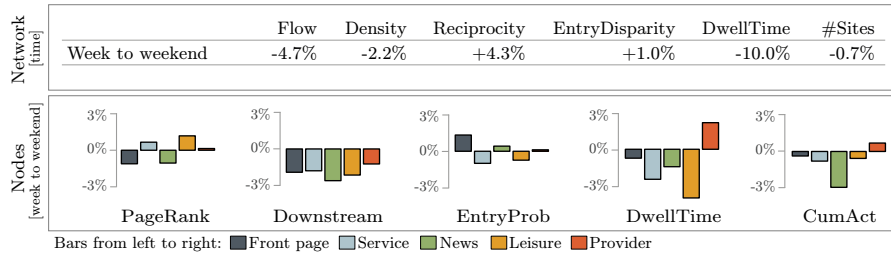


Figure 9.2: Differences between time-based networks: (Top) Network-level metrics. (Bottom) Node-level metrics.

tasks (e.g. checking mails or reading news), 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), 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). This is in accordance to the results of Section 4.5.3.

9.3.3 Returning Traffic

In Chapter 5 we could see that users engage in multitasking during their online sessions, 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 9.3 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. Traffic returning to the same site n_i is represented by an additional edge $w_{i,i}$.

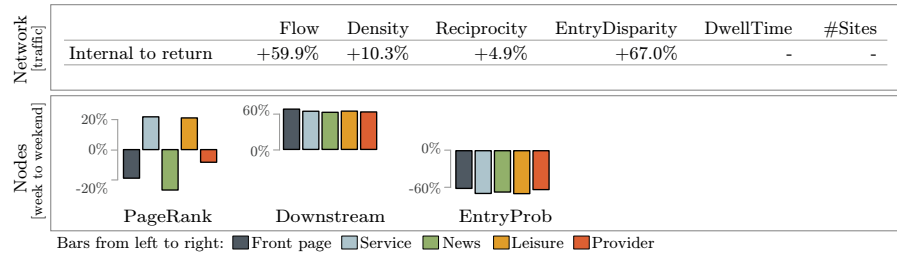


Figure 9.3: Differences between traffic-based networks: (Top) Network-level metrics. (Bottom) Node-level metrics.

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, because 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, since 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 9.5.

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

9.3.4 Upstream Traffic

It was shown [251] that user browsing activity 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 defined in Section 3.3, 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` does not belong to the US network). If the user accessed the network by using a bookmark, or by 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.* `cnm.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 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 [122], since they use bookmarks, remember the domain name and enter

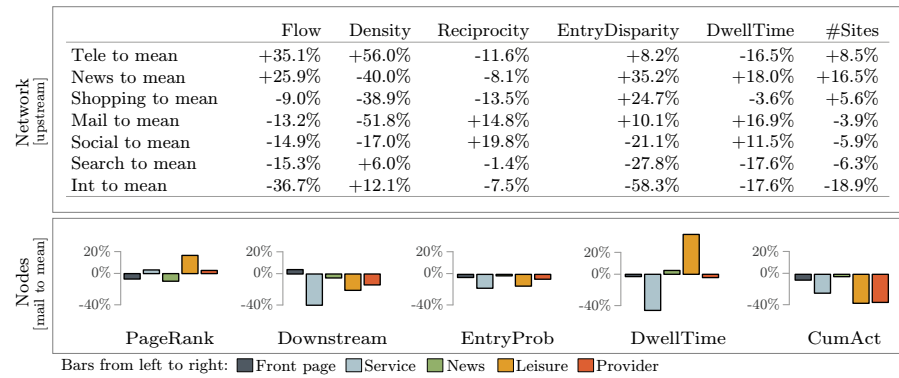


Figure 9.4: Differences between upstream-based networks: (Top) Network-level metrics. (Bottom) Node-level metrics.

it directly, or simply start typing the URL which then get autocompleted. This is also reflected by the network-level metrics in Figure 9.4 (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 6.8 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 9.4 (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 receive 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, since 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 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 this using metrics brought in to measure inter-site engagement.

9.4 The Network Effect

Previous work [168] showed that the popularity of pages depends on the traffic between them, and we already observed in Chapter 6 that there is a strong correlation between site popularity and the inter-site engagement in the network. In this section, we extend the study of Chapter 6 and investigate into the extent of 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 ($US_{01/08}, \dots, US_{31/07}$) and removed networks modelling weekend browsing activity. This is to ensure that our observations are not caused by the difference in browsing activity between weekdays and weekends (see Section 9.3.2).

9.4.1 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 $v_n = (c_1, \dots, c_{261})$, 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 rho coefficient (ρ) between its vectors.

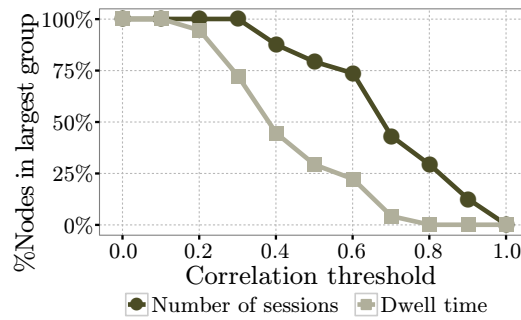


Figure 9.5: The strength of the network effect generated at different correlation thresholds.

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 9.5 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 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 Chapter 4. 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 spend *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.

9.4.2 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 $v_e = (c_1, \dots, c_{261})$, where c_i is the number of clicks on day i , for edge e . We then compare the edges by calculating the Spearman rho coefficient (ρ) between its vectors.¹ Finally, if the correlation is above or below a 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 $a \rightarrow b$ and $a \rightarrow c$, 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 $a \rightarrow c$ and $a \rightarrow d$, we add the site d and the edge $a \rightarrow d$ to that same “network effect pattern”. We continue this process until all edges are compared with each other, and select all patterns that consists 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 $a \rightarrow b$ and $a \rightarrow c$, and $a \rightarrow d$ and $a \rightarrow e$, 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 9.6). 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

¹ 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 similar results.

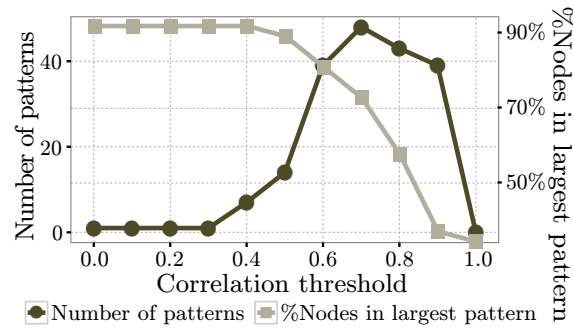


Figure 9.6: Number of network effect patterns and the percentage of sites in the largest pattern generated at different correlation thresholds.

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), because only 1.3% of the correlations are below -0.7.

Examples of Patterns

We focus on the patterns with correlations $|\rho| \geq 0.7$. We divide the patterns in three groups, shown in Figure 9.7. For each, we report the average number of sites, the average reciprocity, and the average transitivity. We recall that the reciprocity [190] describes the probability that the traffic between two sites flow in both directions, whereas the transitivity [190] 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 structure ($Trans = 0$). Simple star-like patterns have one focal site responsible for traffic exchange, whereas complex star-like patterns have more than one

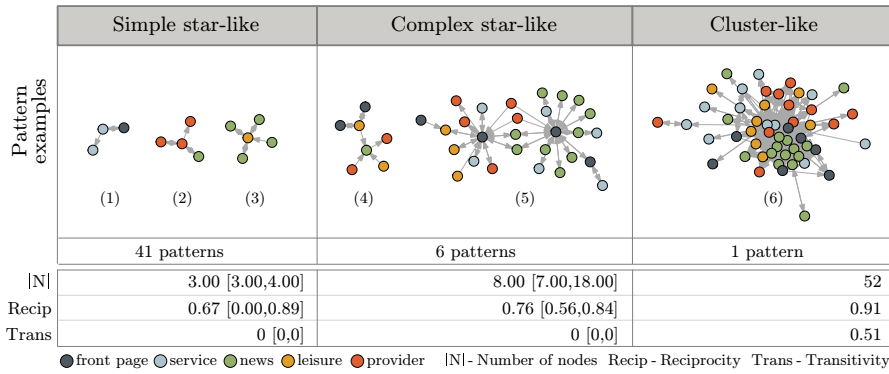


Figure 9.7: Network effect patterns: (1st row) Examples of network effect patterns. (2nd row) Number of patterns belonging to that group. (3rd row) Median of interquartile range of the number of sites ($|N|$), reciprocity ($Recip$), and transitivity ($Trans$) in each group.

focal site. In Figure 9.7 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 hand 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.

9.5 Hyperlink Performance

Previous work [280] demonstrated that hyperlinks help directing users to other sites in a provider network. Motivated by this, we analyse 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, 43K pages were selected. We downloaded the HTML content of the pages, and extracted their hyperlinks.²

We distinguished whether a link 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.

Variations in the Link Structure

We first study whether sites differ in their hyperlink structure. Figure 9.8 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).

Service and news sites have the same amount of on-site and inter-site links. 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%,

² Pages from several sites were not considered, as signing in on the sites was required before downloading the pages.

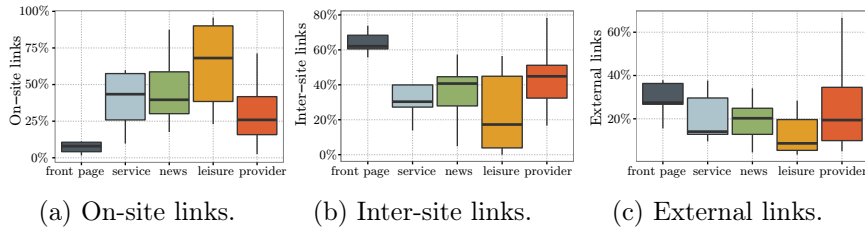


Figure 9.8: Percentage of link types depending on site category.

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, because 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.

Effect of the Link Structure

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

We compare the hyperlink network with the traffic network using the node-level metrics *PageRank* and *Downstream*.³ 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 activity of users when visiting that site. The browsing activity on a site in the traffic network is described by the average percentage of traffic

³ The percentage of external links defines the exit probability.

Table 9.2: Spearman’s rho 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.			(b) On-site, inter-site, and external.			
Hyperlinks	Traffic		Hyperlinks	Traffic		
	PageRank	Downstream		On-site	Inter-site	External
PageRank	0.54	-	On-site	0.54	-0.45	-0.38
Downstream	-	-	Inter-site	-0.40	0.50	-
			External	-	-	0.39

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*).⁴ We then rank sites according to each measure in the traffic and hyperlink network and compare these rankings using the Spearman rho coefficient (p-value < 0.01). The results are presented in Table 9.2.

We observe in Table 9.2a 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).

In Table 9.2b, 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 9.3.3, leaving the network does not necessarily imply less engagement, since users often return to the network within the same session.

⁴ The percentage of external traffic is the same as the node-level metric *ExitProb*.

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 frequently visited 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 [280]. 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.

9.6 Discussion

This chapter studied inter-site engagement in a large provider network. Our network consists of a sample of 73 Yahoo sites based in the United States. We created various network instances to analyse the characteristics of inter-site engagement in provider networks.

In Chapter 4, we observed that site engagement is influenced by the loyalty of users and the day of the week. In this chapter, we observed the same with respect to inter-site engagement. In addition, we analysed how returning traffic (users leaving the network but returning within the same session), and upstream traffic affect inter-site engagement. We saw that leaving the provider network does not necessarily entail less engagement, since many users return later on. As already observed in Chapter 5, 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 also extended the study of the *network effect* in Chapter 6, that is, we investigated the dependencies between site engagement and traffic between sites in the provider network. 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, to not exceed their available time.

Finally, we compared the traffic and hyperlink network with each other, and showed that hyperlinks can influence user browsing activity 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 with sites (users spend less time on sites). This means that site and inter-site engagement influence each other. Improving both at the same time may be difficult.

Limitations. Our work comes with certain limitations. First, we looked at the effect of several dimensions on inter-site engagement separately. It is possible that some dimensions overlap (*e.g.* loyal users drive returning traffic), and that combining these dimensions would provide us with further insights (*e.g.* behaviour of loyal users for each upstream traffic type). However, we wanted to analyse the dimensions separately to obtain clear and focused insights about the effect of each dimension.

Second, although we could show that there is a strong network effect with respect to site popularity, we did not attempt to identify the sources of such effect. It is possible that the popularity of sites decreases or increases simultaneously, or that certain sites (*e.g.* front pages) initiate the spread of traffic through the network, comparable with information diffusion in online social networks.

Finally, we did not consider the position and style of hyperlinks. It is apparent that hyperlinks at the bottom of the pages, and less visible hyperlinks referring to, for instance, the “About” or “Contact” page, are less frequently used. However, we still obtained valuable correlations. We speculate that, when accounting for these aspects, the correlations will become even stronger.



Story-Focused Reading across News Sites

In this chapter, we study story-focused news reading, which occurs when users read several articles related to a particular news development. Our aim is to understand the effect of story-focused reading on site and inter-site engagement and how news sites can support this phenomenon.

10.1 Introduction

The Web has totally changed the news landscape, causing a significant drop in newspaper and radio audiences, and becoming the second source of news in the US, after television [37]. Online news reading is one of the most common activities of Internet users. A survey published by the Pew Research Center in 2012 [37] reported that 39% of news readers get their news online. Users may have different motivations to visit a news site. Some users want to remain informed about a specific news story they are following, such as a big sport tournament (*e.g.* the Baseball World Series) or an important political issue (*e.g.* *Obamacare*). Others visit news portals to read about breaking news and remain informed about current events in general.

While reading news, users sometimes become interested in a particular news item they just read, and want to find out more about it. They may do so to obtain various angles on the story, for example to overcome media bias [225], or to confirm the veracity of what they are reading. Indeed, a study from the New York Times [209] reported that many users still visit established news outlets to confirm a story, no matter from which source

the information initially came from. In this chapter, we study this type of reading behaviour. We describe *story-focused news reading*, or simply *story-focused reading*, which occurs when users read multiple articles about a particular news development or event. In this chapter, *article* refers to a single document, and *story* refers to a set of related articles.

The Web has totally changed the news landscape, because users have the possibility to read news from diverse news sites and other sources. A recent study found that 57% of users routinely get their news from between two to five news sites [36]. Although users increasingly use social media sites to share news they read and find worth sharing, search engines continue to be the most important tool for users to look for articles on news of interest to them; more than 33% of users use search engines regularly to find news [37]. Several search engines offer *news verticals* specifically designed for users to search for news published by online news sites [12; 174].

All this implies that studying news reading *at site level* leads to a rather limited view about user reading behaviour, since users engage with several news providers and employ various means to find the articles they are interested in. We therefore consider inter-site engagement in this work, by analysing how users engage with news reading across news providers and we study what other sites they use to access the news articles they are interested in.

Our study is based on a large sample of user interaction data on 65 popular news sites publishing articles in the English language. We analysed 4.9M news reading sessions covering a total of 2,536 stories comprising 25,703 news articles. Stories range from policy issues such as the threat of the US government shutdown (October 1 to 16, 2013), and the NSA spying scandal, to less weighty issues such as the Draconid meteor shower, and specific sport events.

We study the characteristics of story-focused news reading, and found that users spend more time reading and visit a larger number of news sites when focusing on a story. We also found that story-focused reading leads to a higher traffic flow between the news sites. This implies that inter-site engagement is more predominant when users focus on a story.

When asked about news reading online, many users of news sites have said that links to related information on a news article page are important [36]. News sites recognise that users want to further inform themselves, and provide information on different aspects or components of a story they are covering. They also link to other articles published by them, and sometimes even to articles published by other news sites or sources.

We also analyse how news sites promote story-focused news reading, by looking at how they link their articles to related content published by them, or by other sources. For instance, we demonstrate that providing too few links to related content may lose an opportunity to keep users on a news website, but that too many links can have detrimental effect on users' reading experience.

We finally show that promoting story-focused news reading leads to a higher engagement of the users with the news site (*i.e.* higher site engagement) – with respect to the reading time, and with respect to the time it takes for users to return to a news site. Interestingly, we also demonstrate that linking to external sources, *i.e.* promoting inter-site engagement, does not have a negative effect on the engagement with the news site – it does not influence the reading time, and it has even a positive impact on the time it takes for users to return to the news site.

The chapter is organised as follows. We start by covering previous work in Section 10.2. In Section 10.4 we show that story-focused reading exists and that it is not a trivial phenomenon. Section 10.5 studies which users focus on stories and on which topics. Story-focused reading sessions are characterised and compared with non-story-focused reading sessions in Section 10.6. We analyse how news providers promote story-focused reading by offering links to related content and which effect it has on the engagement with the provider in Sections 10.7 and 10.8, respectively. We discuss our main findings in Section 10.9.

10.2 Related Work

Reading behaviour. How users browse the Web has been studied in many contexts. In the online industry, knowing how users interact with a site is used to assess users' depth of engagement with the site, employing metrics such as time spent on a site and return rates. Studies focusing on user reading behaviour on news sites have shown, for instance, how users engage with news sites [194] and that their interests are location-dependent and change over time [166]. These findings have helped in developing personalised news recommendation systems [57; 160]. We also study user news reading behaviour, but with respect to users reading multiple articles related to a same story, and how this affects the engagement with the news provider site.

Users often visit more than one site during their online sessions. Already in 1995, Catledge *et al.* [35] showed that while most sessions involved a single website, some sessions involve many websites. Our work regarding the multitasking behaviour of users in Chapter 5 has validated this. The fact that users access and engage with several sites may indicate a task-oriented navigation or task-focused searching in the context of search [213]; for instance, to plan for a trip, several sites are accessed to compare locations and hotels. In this work, we look at this in the context of users engaging with several news sites while reading news articles about a story.

Online news reading has also become more interactive [38; 119]: news articles contain multimedia content (pictures, videos, audio clips), social features (commenting, sharing), and hyperlinks embedded into news articles. This has been shown to increase user engagement, with users having more control over the content they wish to consume, and how they consume it [201]. In this work, we look at the effect of some of these on story-focused reading.

News stories. Although many studies analysing the online behaviour of users while reading news exist, little is known about how users consume news articles related to a specific story. Topic detection and tracking (TDT) has been a research topic for many years; the aim is to group articles that refer to the same news story or topic [27; 262]. Other works analyse the evolution of stories over time [103; 158] or identify chains of news articles that show how two stories are connected with each other [230].

These works provide the means for news sites to provide, if they wish, overviews about the stories they are publishing.¹ However, individual news sites typically do not promote story-related reading, in contrast with online tools performing news aggregation, such as the Europe Media Monitor² and Storify.³ In this chapter, we study the reading behaviour of users around specific stories, demonstrate that story-related reading happens, and show its effect on user's daily consumption of news.

Link economy. Hyperlinks are a simple but powerful tool on the Web to build connections between content and direct users to specific websites or web pages. Although news aggregators such as Google news are sometimes accused of “stealing” audience from traditional news outlets, they direct significant traffic to news sites, because they are widely used by users to search for information [63].

¹<http://newspulse.cnn.com/>

²<http://emm.newsbrief.eu/>

³<https://storify.com/>

A news provider linking strategy depends strongly on the type of content it provides. Whereas blogs rely on hyperlinks to be reached by users and to direct users to websites [69], most news sites do not link to each other, because of competition. Instead, they invest time and effort into connecting their news articles with each other using hyperlinks [38]. This is to keep users on their site and increase user engagement [23; 255].

Recent studies [61; 63; 220] advocate that the right linking strategy, such as providing hyperlinks between news sites to provide users with more information about a story, can actually increase profits in a costless way. In addition, it provides a more interactive, credible, transparent, and diverse news reading experience to users.

Finally, McCreddie *et al.* [174] showed that news articles are not the only important information source satisfying news-related queries in search engines. When searching for news, users like to see Wikipedia pages, blog posts, and tweets. This is sometimes sufficient to satisfy their information needs. Linking news articles to other information sources [38; 182; 227] allows users to learn about the story context, *e.g.* the background and history of the story, and opinions and discussions around it. A number of news sites have already recognised this, and provide links to articles published by other news providers or to sites containing background information.⁴

In this work, we study the different linking strategies of news sites and their impact on user news reading behaviour.

10.3 Dataset

Our dataset is based on one month (October 2013) of anonymised interaction data from a random sample of users who gave their consent to provide data through the Yahoo Toolbar. This sample consists of 800K users, and 325M page views.

We considered the 100 most visited English-speaking news sites according to the ranking provided by Alexa.⁵ Alexa's ranking includes both traditional news outlets (*e.g.* The New York Times or CNN) and news aggregators (*e.g.* Yahoo News). We selected news providers based in the US, UK, Canada or Australia; this resulted in 65 news sites listed in the Appendix in Table A.4. We also consider news providers that cover only specific gen-

⁴<http://www.nytimes.com/2006/07/31/technology/31ecom.html>

⁵<http://www.alexa.com/topsites/category/Top/News>

res (referred to as section); *e.g.* bankrate.com and dailyfinance.com report mainly about investment and financial stories.

To ensure that no strong bias in the browsing data affect our results and their applicability, we compared the Alexa ranking with our data on the basis of total user traffic, and found that the two correlate well (Kendall's $\tau = 0.62$, and Spearman's $\rho = 0.80$). Hence, the insights gained in this research are not specific to Yahoo toolbar users.

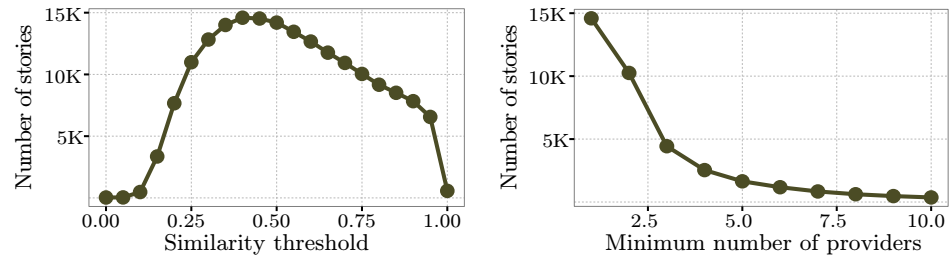
Articles and stories. A *news article* or simply *article* is a single document on a news website (an HTML page). For each of the 65 sites, we used various rules, based on regular expressions, to distinguish between visits to an actual news article from visits to other parts of a news website, such as its homepage or section pages (*e.g.* “politics” or “world news”). Articles visited by less than 5 users during a day, which correspond mostly to very old articles, were not considered for this study. This removed 8.9% of the browsing events, and lefts us with 98,241 news articles.

A *news story* or simply *story* is a collection of articles related to the same news event. To identify whether two news articles belong to the same story, we applied the cosine similarity metric on the two article texts [15], where a value above a given threshold means that the two articles are related to the same story. We experimented with several threshold values, as shown in Figure 10.1a. For 0.4, we reach the maximum number of stories. We therefore use this value as our threshold.⁶

We also removed *niche stories*, which are stories covered by very few news providers; these stories are likely to be region-specific. To identify niche stories in our dataset, we calculated the number of stories identified based on the minimum number of news providers we considered. We experiment with several numbers, and plot the outcomes in Figure 10.1b. We see that many stories are published by only few providers. The number of stories decreases fast as the minimum number of considered providers increases, and then slows down at around 3-4 providers. As a conservative setting, we define a story to be *niche* if it is covered by 3 or less providers, and *top* if it is covered by 4 or more providers.

This process results in 2,536 top stories, about 82 per day, and 25,703 articles. On average, each of these stories has 14 articles (median 8), and is

⁶Using a threshold of 0.3 or 0.5 yields a similar number of stories. Note that a story must involve at least two articles.



(a) Number of stories generated at different similarity thresholds.

(b) Number of stories depending on minimum number of providers.

Figure 10.1: Selecting the threshold to decide when two articles relate to the same story, and the threshold to decide when a story is niche.

covered by 7 providers (median 5). The distribution of visits to stories is very skewed, with an average of 2,482 users per story (median 758). The number of visits correlates moderately with the number of articles about the story (Spearman’s $\rho = 0.67$) and with the number of providers that cover it ($\rho = 0.54$).

Sections. We also used several regular expressions to extract from each article’s URL, the section it belongs to. We noticed, for example, that a story can include an article listed as “politics” in one site and a second article under “world news” in another site. We therefore assign one section per story, which we chose to be the most common one under which its associated articles have been published. We do this only if the selected section contains 50% or more of the articles.

We were able to assign a section for 92% of the stories in our dataset: 22% business, 20% life/entertainment, 23% local, 6% science, 9% sports, and 12% world news stories. The remaining stories (8%) are labelled as “misc.” (e.g. the story about the “FIFA World Cup in Qatar” belongs to the sport and world section).

News reading sessions. A *news reading session* is an online session⁷ in which *at least one* news article of the selected stories is accessed. In our dataset, we extracted a total of 4.9M news reading sessions.

In this chapter, we also analyse how users navigate to news articles (Section 10.6.3), and to which pages or sites they are navigating to afterwards

⁷See Section 3.2 for a detailed description of online sessions.

(Section 10.7). Therefore, we distinguish whether the user came from (navigated to) a page of the same provider (*internal traffic*), or from (to) somewhere else on the Web (*external traffic*).

For both cases, we also differentiate the traffic from (to) articles ([Internal/External Article]) or other pages ([Internal/External Non-Article]) of the provider sites in our list. We annotate the remaining (all [External]) sites using the schema described in Section 3.3 as follows:⁸

- 1022 further news sites and blogs [News Non-Top]
- 42 news aggregators and online RSS feeds (*e.g.* Google news, FriendFeed) [News Aggregator]
- 39 social media sites (*e.g.* Twitter) [Social Media]
- 5 mail sites [External Mail]
- 25 multimedia sites (*e.g.* YouTube) [Multimedia]
- 52 reference sites (*e.g.* Wikipedia) [Reference]
- 10 search engines (*e.g.* Google, Bing) [Search]
- 812 organisation sites (*e.g.* nasa.gov) [Organisation]
- 7 front pages (*e.g.* AOL) [Front page]
- 17K uncategorised sites [Other]

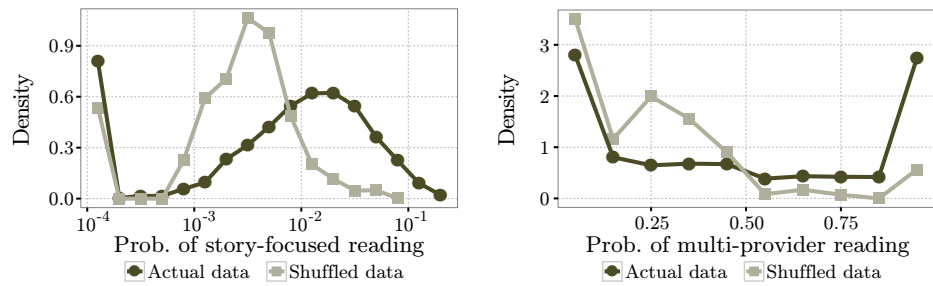
10.4 Story-Focused News Reading

Now, we show that story-focused news reading exists and that it is not a trivial phenomenon, *i.e.* not merely a consequence of how articles are distributed among stories.

10.4.1 Shuffle Test

We first determine whether story-focused reading occurs because many articles belong to a story, or because users are interested in reading articles related to a specific story. To answer this question we perform a *shuffle test* (similar to [7]). We create an alternative dataset of news reading sessions

⁸ Depending on whether we analyse the upstream or downstream traffic, we made the following simplifications. In Section 10.6.3, the multimedia, reference, and organisation sites are denominated as [Ext. Other]. In Section 10.7, news aggregator sites are merged to [Ext. News Non-Top], and mail and search sites, and front pages are part of [Ext. Other].



(a) Distribution of the probability of story-focused reading. (b) Distribution of the probability of multi-provider reading.

Figure 10.2: Shuffle test showing the difference in story-focused reading and multi-provider reading between the actual dataset and the shuffled dataset.

that has the same distribution of session length, but with random articles in them. We call this alternative dataset the *shuffled dataset*.

We next calculate the *probability of story-focused reading* for a given story s :

$$\frac{\text{\#story-focused sessions of } s}{\text{\#sessions including } s}$$

where the story-focused sessions for a story s are those in which a user visits two or more articles related to the story s . The distribution of this probability across all stories is shown in Figure 10.2a.⁹ We also calculate the *probability of multi-provider reading* for a given story s :

$$\frac{\text{\#story-focused sessions of } s \text{ in two or more providers}}{\text{\#story-focused sessions of } s}$$

which is shown in Figure 10.2b.

We observe a clear difference between the actual dataset and the shuffled one. The probability of story focused reading is about 4 times larger than of the actual data (0.019 vs. 0.005), and the probability of multi-provider reading is about two times larger than of the actual dataset (0.48 vs. 0.25). A Kolmogorov-Smirnov (K-S test) test confirms that the difference between the distributions is statistically significant ($p < 0.01$). This indicates that story-focused reading is observed due to users deciding to read multiple articles associated with a story.

⁹In this and other log-scale plots, we added to each value a small constant (0.0001) to represent zeros in the log scale.

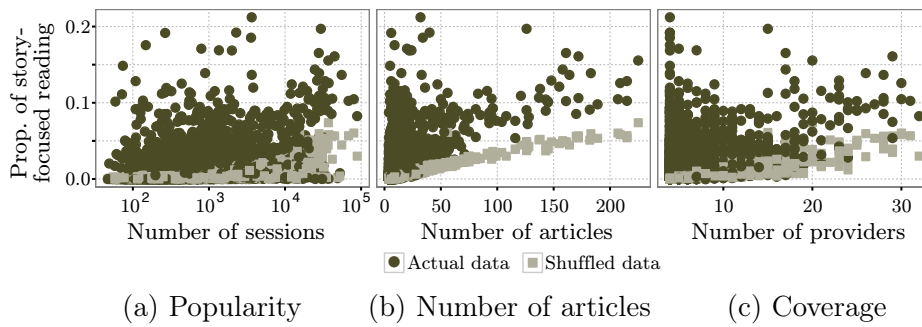


Figure 10.3: Story-focused reading and popularity, number of articles, and coverage.

10.4.2 Popularity and Providers

We analyse whether story-focused reading depends on the popularity of a story, on its number of articles, or on the number of news providers that cover it. The *popularity* of a news story is defined as the number of sessions where users have read articles related to that story.¹⁰ We again compare our dataset with the shuffled dataset.

In Figure 10.3(a) we plot the *probability of story-focused reading* based on the story popularity. We observe that story-focused reading is not necessarily related to popularity. Even stories that are not popular engage users in story-focused reading. The probability of story-focused reading given its popularity is lower than what is observed with the shuffled dataset (Spearman’s $\rho = 0.30$ vs. $\rho = 0.57$ in the shuffled dataset). Overall, the probability of story-focused reading is comparable across all levels of popularity; this indicates that personal interests trigger users into story-focused reading.

Story-focused reading is also not merely a consequence of having a story reported through many articles. Figure 10.3(b) shows the probability of story-focused reading as a function of the number of articles published about that story. We observe that even stories having few articles written about them engage users in story-focused reading. Compared with the shuffled dataset, the correlation between this probability and the number of articles is lower in the actual dataset (Spearman’s $\rho = 0.61$, vs. 0.80 in the shuffled dataset). For instance, in our dataset we obtain a probability of story-focused reading equal to 0.1 for the two stories “Royal Christening of Prince George” (Octo-

¹⁰The traffic volume has been scaled with an arbitrary but constant factor for confidentiality.

ber 23, 2013) and “Draconid Meteor Shower” (October 7, 2013). The former story has 95 articles associated with it, whereas the latter story has only 8 articles.

We reach similar conclusions when relating the probability of story-focused reading to how many news providers are reporting about a story, as shown in Figure 10.3(c). The fact that several news sites report the same news story is not what promotes story-focused reading. The correlation is lower than in the shuffled dataset (Spearman’s $\rho = 0.36$ vs. 0.62 in the shuffled data).

Overall, we can conclude that story-focused reading is a real phenomenon and not simply a consequence that some stories are more read, have more articles written about them, or are covered by more providers.

10.5 Users and News Sections

Now, we study which users are more likely to engage in story-focused reading, and on what type of stories.

10.5.1 Story-Focused Reading and Users

The percentage of users that engage, at least once, in story-focused reading is 16% in our one-month dataset. As expected, avid news readers are more likely to engage in story-focused reading: 64% of the users with at least 15 reading sessions in our one-month dataset have at least one story-focused session. However, this does not imply that the more articles a user is reading, the more often s/he is engaging in story-focused reading. In fact, the correlation between these two variables is only $\rho = 0.45$.

We investigate further into this. Figure 10.4 shows two distributions, one for number of sessions and the second for number of story-focused sessions, per user. The distribution of the number of sessions per user can be fitted with a Zipf-Mandelbrot distribution (following [6]); this is a highly-skewed distribution, but with more users having a mid-range number of sessions than a power-law would predict. The number of story-focused sessions per user follows a power-law (Pareto) distribution with exponent 2.7. This means that sessions with story-focused reading are different from normal web sessions—the browsing characteristics are different. Indeed, although many users have a mid-range reading activity, most of them only have a few story-focused sessions, whereas few users have many more.

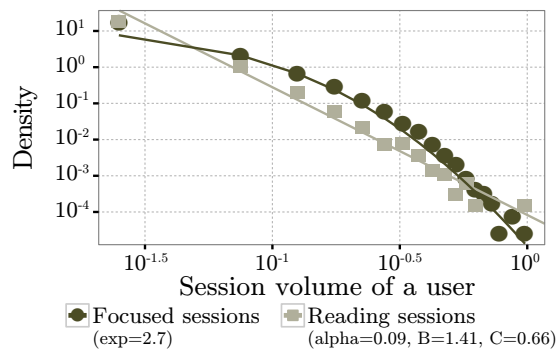


Figure 10.4: Distribution of reading and story-focused sessions over the users on a log-log scale. The x-axis has been normalised so that the maximum is 1 (10^1).

10.5.2 Story-Focused Reading and Sections

We study whether users are more likely to engage in story-focused reading in some sections than others. Table 10.1 reports various statistics for each section. The six sections under consideration were described in Section 10.3.

Stories in the sport and business sections are the most likely to engage users in story-focused reading. The percentage of users with story-focused sessions is 4.4% and 3.7%, respectively. An example of a story that appeared prominently in the business section was the threat of government shutdown in the US (from October 1 to 16, 2013). Many sport stories are about football or baseball games, such as the game between the Tigers and the Red Sox (October 19, 2013). The articles in these sections are also relatively long on average (714 and 680 words, respectively). This suggests that story-focused reading is probably not caused by stories having articles that are too short; indeed, we do not observe a correlation between the average length of the articles on a story and the probability that users will engage in story-focused reading with that story (Spearman's $\rho = 0.24$).

Interestingly, news stories related to sport are not very popular (14% of users) compared to stories of other sections such as business and local (25% and 20% of users, respectively), showing again that story-focused reading is not driven by the popularity of stories. We speculate that story-focused reading depends on location (*e.g.* *Obamacare* may be more relevant to people living in the US) and personal interests (*e.g.* users that are keen on sports may read more sport news). The same was observed in [166] in the context

Table 10.1: Sections in which it is more likely to find story-focused reading. We provide the percentage of users that focuses on at least one story on that section with respect to all users on that section, the percentage of users on that section, the average article length (in words), and a story example.

Section	%Foc. users	% Users	Art. length	Example
sport	4.4%	14%	714	Tigers vs. Red Sox game
business	3.7%	25%	680	US government shutdown
world	2.3%	9%	624	NSA tapped Merkel's phone
life/ent.	2.1%	20%	454	Royal prince christening
misc	2.1%	5%	577	FIFA World Cup in Qatar
local	1.9%	20%	540	BART strike / San Francisco
science	1.8%	4%	578	Draconid meteor shower

of standard news reading. In addition, the interests of users change over time, and spikes of interest can be observed around large events such as the FIFA World Cup or the elections of the European Parliament.

Overall, avid users are more likely to engage in story-focused reading. However, personal interests and the local-context matter, whereas article length does not.

10.6 Session Characteristics

We analyse *story-focused news reading sessions*, or simply story-focused sessions, which are news reading sessions where users visit at least *two* news articles related to a same story.¹¹ Since we are studying user reading behaviour across news sites, we first define a network of news sites and then employ the inter-site engagement metrics of Section 6.5.

Networks and metrics. We used the selected 65 news sites to extract a number of weighted directed network instances, where the set of nodes corresponds to the news sites and the set of edges to the user traffic between them. An edge $n_i \rightarrow n_j$ is created when a user read an article on site n_i and then an article on site n_j . We do not consider whether users visited other sites while navigating from site n_i to n_j . In Sections 10.6.1 and 10.6.2, we are only interested in the traffic flow between news sites, no matter what other sites (*e.g.* search) are responsible for this flow. How users reach the articles is analysed in Section 10.6.3.

¹¹ We note that 75% of the sessions contain only one news article, and by definition cannot be story-focused.

The edge weight $w_{i,j}$ represents the size of the user traffic between two sites, which depends on the selected sessions we want to analyse and compare. For instance, in Section 10.6.2, we analyse user reading behaviour depending on the depth of story-focused reading, that is, the number of articles read about a story. For this purpose, we created 7 network instances using the story-focused sessions of depth 2, ..., 7, and > 7 .

The metrics used in this section are defined in Chapter 6. We focused on metrics at network-level, and calculated the average time users spend in the network (*DwellTimeS*), and the number of news sites visited during the reading sessions (*#Sites*). In this chapter, we refer to these metrics as *Duration* and *#Providers*, respectively. We also measured the traffic flow (*Flow*) and entry disparity (*EntryDisparity*) in the network. The entry disparity reflects whether users start their reading sessions from a smaller subset of news sites (high *EntryDisparity*) or whether the in-going traffic to the network is equally distributed over the news sites (low *EntryDisparity*). In case the users only visit one news site per reading session, *i.e.* there is no traffic between the nodes in the network, the metric simply reflects the differences in popularity between the news sites. Whether there is traffic in the network is reflected by the traffic flow metric (*Flow*), which measures the extent to which users navigate between the news sites.

10.6.1 Focused versus Non-focused Sessions

Story-focused sessions have several characteristics that distinguish them from non-story-focused sessions (sessions where no story-focused reading is observed). Figure 10.5 compares the two, grouping them by session length (number of articles visited in a session). Using a K-S test, we can confirm that the described differences are statistically significant ($p\text{-value} < 0.05$).

Figure 10.5 (a) shows that it is more likely to have shorter reading sessions: 67% of the reading sessions contain only two news articles whereas for 3% of the reading sessions the users read 5 news articles. We observe that when the session length increases (more articles are read), the probability that a session is story-focused increases (Figure 10.5 (d)). For instance, 41% of the sessions with 4 articles are story-focused sessions. This shows that story-focused reading becomes more predominant the longer the user spends time reading news.

We also see that users spend more time in their news reading activity when focusing on a specific story, compared to when they access articles about different stories (Figure 10.5 (b)). Story-focused sessions are at least 15%

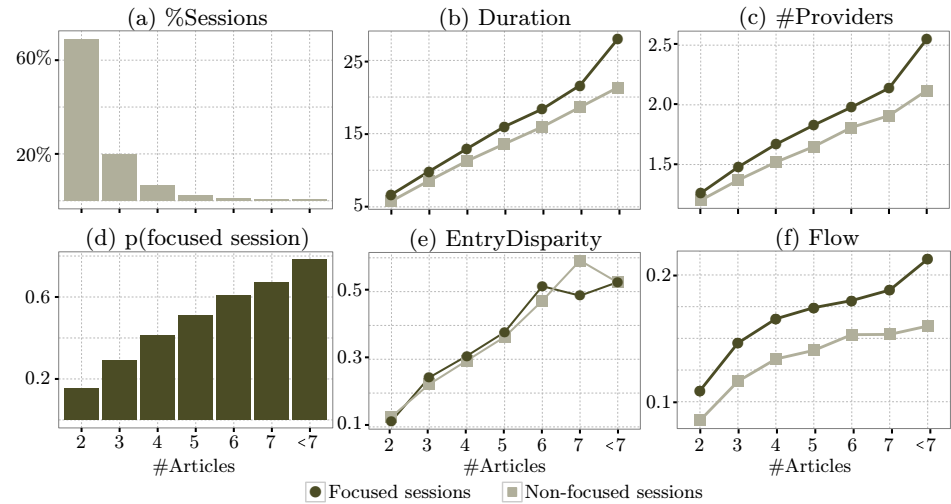


Figure 10.5: Comparison of story-focused and non-story-focused sessions at different session lengths (#Articles), including the percentage of reading sessions at each length (a), the probability that a session is story-focused (d), the total duration of the session in minutes (b), the number of distinct news providers visited (c), the entry disparity (e), and the traffic flow (f).

longer, and the difference increases with the session length. For instance, for sessions with 4 articles, the session durations are on average 11.24 minutes (in non-focused sessions) and 12.92 minutes (in focused sessions).

We can report a similar trend when looking at the number of news providers visited (Figure 10.5 (c)). For sessions with 4 articles, the average number of visited news providers is 1.52 (non-focused sessions) and 1.67 (focused sessions). The difference is even higher when looking at the traffic flow (Figure 10.5 (f)), showing that users more often navigate between news sites during story-focused sessions. Overall, the results suggest that inter-site engagement is more predominant in story-focused sessions.

Finally, we do not observe a difference in the entry disparity between story-focused and non-story-focused sessions (Figure 10.5 (e)). For both types of sessions the entry disparity increases in the same way, suggesting that the longer the reading sessions the more frequently users start their sessions with a subset of news providers.

Table 10.2: Statistics of story-focused sessions of different depth (in-story articles), including percentage of sessions, number of in-story and out-story articles, duration in minutes, and number of distinct providers.

Sess. %	Number of articles		Duration		#Prov.	Flow	EntryDisp.
	In-story	Out-story	Total	Per-article			
85.03	2	0.89	6.67	3.34	1.27	0.22	0.01
11.48	3	1.09	10.48	3.49	1.53	0.38	0.02
2.43	4	1.45	14.29	3.57	1.79	0.50	0.04
0.69	5	1.67	18.23	3.65	2.05	0.58	0.04
0.23	6	1.80	20.09	3.35	2.31	0.62	0.15
0.08	7	2.67	23.09	3.30	2.36	0.66	0.15
0.06	>7	3.05	25.03	2.79	3.19	0.73	0.25

10.6.2 Depth of Story-focused Reading

Session length only takes into account the number of articles visited during the session. These articles may not necessarily relate to a same story. We study now the reading behaviour depending on how many articles of the same story are accessed (number of in-story articles), called the *story depth*. We also report the number of out-story articles (articles that do not belong to the story the user is focusing on), and various other statistics (see Table 10.2). In cases where the user is focusing on several stories within a session, we calculated the browsing activity with respect to each story. We note, however, that only 2.36% of the story-focused sessions have the user reading about more than one story.

Deeper story-focused sessions are naturally longer. They also involve a larger number of news providers, and the traffic flow is higher. The high entry disparity indicates that some news sites are more often used to start longer story-focused session than others. We speculate that these news sites promote story-focused reading more than the other news sites. We return to this in Section 10.8. The number of out-story articles is higher as the session depth increases; however, the in-story articles always constitute the majority of articles read in story-focused sessions.

In sessions with 5 or less in-story articles (99.6% of the sessions), we see an increase in the *per-article dwell time*: users spend time reading the individual articles they are accessing. For the 0.4% of sessions with 6 in-story articles or more, the dwell time decreases. This is in accordance with results reported in Section 5.5.1 about user multitasking behaviour, and suggests that users are skimming the articles, probably because the articles contain increasingly more redundant information.

To verify this, we measure for each article read in a session the number of words (information) that do not occur in any other article of that session. In sessions with 2 in-story articles, each article contains on average 49% unique words, whereas in sessions with 7 or more in-story articles, the percentage of unique words per article is on average 33%. This suggests that with each article read in a session, users skip increasingly more parts of it (*e.g.* information known from previous articles), and skim the article for specific or new facts.

10.6.3 Upstream Traffic

We want to understand how users reach the articles when they actually engage in story-focused reading. Using the HTTP referrers available in our browsing dataset and the categorisation schema of Section 10.3, we study which sites users navigate from when engaging in story-focused reading. Since we want to analyse how users find story-related articles after they decided to focus on a story (*i.e.* after the first article was read), we consider *upstream traffic* in relation to the second, third, *etc.*, article being accessed. Table 10.3 shows the percentage of *upstream traffic* for the considered site categories, grouped by story depth (*i.e.* number of in-story articles).

Most of the traffic to an article comes from other pages of the same news provider. In sessions with 2 in-story articles, 78.8% of the traffic is coming from another page of the same provider (*internal traffic*), and only 21.2% of the traffic originates from somewhere else on the Web (*external traffic*). However, the dominance of internal traffic decreases as more articles are read. For example, if users read more than 7 articles about the same story, only 55.1% of the traffic comes from another page of the same provider.

Interesting is that the internal traffic is mainly driven from non-article pages of the news provider. Looking again at story-focused sessions with depth 2, in only 17.8% of the article views the users navigated from another article of the news provider; in 61.0% of the article views the users employ other means on the provider site to access related articles. With respect to this, we can report that for on average 57% (median=67%) of the article views per news provider, the user clicked on a link on the front page of that provider. This suggests that the linking strategy of many providers does not support story-focused reading *at* the article level, since users are more likely to return to the front page to search for another article related to the story.

We look now at the external upstream traffic, and discuss in particular the values obtained for sessions with more than 7 in-story articles. The longer the story-focused sessions, the more articles are accessed from webmail sites and other sources (3.8%, and 11.6%, respectively). The same applies for less popular news sites and social media sites. We see that 11.2% of the upstream traffic comes from less popular news sites (“News Non-Top”). This shows that in the context of story-focused reading, inter-provider linking can increase the traffic to the most popular news providers. In addition, the usage of social media sites increases (8.4%). This showcases the increasing importance of social media sites as a source of traffic for people interested in having in-depth information about a story. For instance, Twitter allows users to click on a hashtag or search for it (*e.g.* #Obamacare), and then access multiple related articles. We also see that front pages are frequently used to access related articles. Although the traffic coming from the front pages decreases as the story depth increases, it increases again for sessions with more than 7 in-story articles (5.7%).

For all other types of sites, we observe that the traffic increases first (until around 5 in-story articles), then the traffic decreases as the story depth increases. In the previous section, we observed the same behaviour for the dwell time per article. This suggests that when users are skimming many articles (because they are redundant, or to search for a specific piece of information), these upstream traffic categories are less frequent.

Upstream traffic and inter-site engagement. Finally, we characterise inter-site engagement during story-focused sessions depending on the source of upstream traffic. Following Section 9.3.4, we define network instances depending on the upstream traffic in relation to the first article being accessed (*i.e.* when entering the network). We do not consider the upstream traffic of the second, third, *etc.*, article, because we want to analyse the differences in story-focused sessions depending on where the user initially came from. Table 10.3 reports the traffic flow and entry disparity for the upstream-based networks (column “Flow” and “EntryDisp.”).

We observe that the flow in the network is much lower when the traffic originates from the same news provider compared to traffic that comes from somewhere else on the Web (0.21 and 0.42, respectively). This implies that users read articles from a larger set of news sites when using, for instance, search and social media sites. News providers, on the other hand, try to keep users on the site by providing other news articles related to the story on their site.

Table 10.3: Percentage of upstream traffic sources as a function of the depth of a story-focused session. The traffic is divided into two main categories (internal and external), and for the latter, in seven sub-categories. The last two columns report the flow and entry disparity of the networks defined by each upstream traffic source.

	Number of in-story articles							Flow	EntryDisp.
	2	3	4	5	6	7	>7		
Internal	78.8	77.2	75.5	73.2	73.3	69.5	55.1	0.21	0.030
Article	17.8	22.9	26.4	28.4	30.2	28.3	16.8	0.20	0.022
Non-Art.	61.0	54.3	49.1	44.8	43.1	41.2	38.3	0.21	0.032
External	21.2	22.8	24.5	26.8	26.7	30.5	44.9	0.42	0.012
Other	5.5	5.4	6.7	7.3	8.0	9.4	11.6	0.31	0.019
News Non-Top	1.5	1.9	2.1	2.4	2.2	2.5	11.2	0.34	0.032
Social media	1.2	1.0	1.3	1.7	2.0	3.1	8.4	0.35	0.021
Mail	1.3	1.8	2.1	2.5	2.7	3.6	3.8	0.21	0.032
Front page	5.3	5.0	4.0	4.2	3.7	3.8	5.7	0.32	0.066
Search	3.7	4.4	5.1	5.3	5.0	5.2	2.8	0.49	0.016
News Article	0.6	0.8	0.9	1.0	0.9	0.8	0.6	0.28	0.076
News Non-Art.	0.8	1.1	0.9	1.0	1.0	1.0	0.5	0.44	0.025
News Aggr.	1.3	1.4	1.4	1.4	1.2	1.1	0.3	0.50	0.030

We discuss now the networks based on external upstream traffic. The networks that are formed by the traffic coming from external front pages and external news article pages have a low traffic flow (0.32 and 0.28, respectively), and the highest entry disparity (0.066 and 0.076, respectively) compared to the other networks. We speculate that these pages mainly provide links to a subset of providers in the network. Therefore, it is more likely that users visit these providers (high *EntryDisparity*), and they also prefer to stay on the provider site when focusing on a story (low *Flow*).

In contrast, if the upstream traffic originates from news aggregators, search or social media sites, the corresponding networks are characterised by a high traffic flow (0.50, 0.49 and 0.35, respectively). This suggests that if users use the mentioned upstream traffic categories, they access many different providers and often switch between them during story-focused sessions. Interestingly, the entry disparity of the networks formed by social media (0.021) and search sites (0.016) is much lower than the entry disparity of the network formed by news aggregators (0.30). This shows that if social media sites and search engines are used to access story-related news articles the traffic gets equally distributed over the news sites in the network, whereas news aggregators favour certain news providers by directing more traffic to them.

In this section, we showed that story-focused sessions differ from non-story-focused sessions – users spend more time when focusing on a story and the inter-site engagement is higher. We also observed that story-focused sessions differ depending on the depth of story-focused reading, which means users employ to access the articles related to a story.

Finally, we showed that the support for story-focused reading on the article pages of news providers has substantial room for improvement, since users employ other means to access related articles (*e.g.* the front page of the provider or social media sites). Taking into account that story-focused reading leads to a higher reading time, news providers might underestimate such an opportunity to keep users engaged with the site. We return to this in Section 10.8.

10.7 Hyperlink Performance

We study the different linking strategies used by news providers, and how (and whether) these strategies influence story-focused reading. Our intuition is that by offering links to related content, news providers encourage users to engage in story-focused reading.

The first step is to investigate how such links perform (do they get clicked?), which is the aim of this section. In the next section, we study the effect of these links on user engagement (Section 10.8).

Extracting inline links. We download the HTML content of the articles visited by the users in our one-month dataset.¹² We assume that any link within the body of an article, an *inline link*, connects that article to a page that is related to it. A manual inspection of several of the news websites under study shows that inline links point in most cases to articles belonging to the same story. This is a common strategy to provide additional information about a news item and has been shown, when properly deployed, to have a positive effect on users' news reading experience [10]. Inline links can be *internal links*, which point to a page of the same news provider, or *external links*, which point to a page of a different provider.

¹²Articles from the Wall Street Journal were not considered, because those articles could not be freely downloaded.

Although news sites may provide links to related articles in a specific panel (*e.g.* “Related articles”), identifying such cases proved to be complex and introduced some level of ambiguity across providers. Therefore, we do not attempt to distinguish classes of non-inline links. We focus on inline links.

Link popularity and performance. We group inline links based on their type (*e.g.* linking to “Internal News Article” versus “External Multimedia”), and their position in the text. We use two metrics, *popularity* and *performance*, to compare the two inline link groups. Popularity is concerned with how many inline links belong to the group, whereas performance relates to how often the inline links in the group are clicked.

We measure the *popularity* of an inline link group as the percentage of inline links that belong to that group. To measure the *performance* of an inline link group, we calculate the probability that a user clicks on a link of that group using its frequency of occurrence ($\text{propLink} = \frac{\text{\#inline links of that group}}{\text{\#inline links}}$), and compare it with the real click probability ($\text{propClick} = \frac{\text{\#clicks on inline links in that group}}{\text{\#clicks on inline links}}$). The difference indicates the performance of the links:

$$\text{LinkPerf} = \frac{\text{propClick} - \text{propLink}}{\text{propLink}}$$

In our dataset, 75.45% of the article pages have inline links. On average, only 6.4% (4.5% internal and 1.9% external) of the links in an article page are inline links. However, on average 9.4% of the visitors are following these links. Therefore the proportion of traffic over inline links is almost twice as large as their presence in article pages. In the rest of this section, we dive into how these links perform.

10.7.1 Number of Inline Links

Figure 10.6 shows the number of clicks on inline links with respect to the number of inline links in the article. There seems to be a “sweet spot” around 10 inline links per article. The number of clicks increases until reaching about 9 to 11 inline links in the article, and then stagnates. Clicks per-link also start to drop around that same number, and articles having more than 29 inline links tend to elicit fewer clicks on inline links than articles with less inline links.

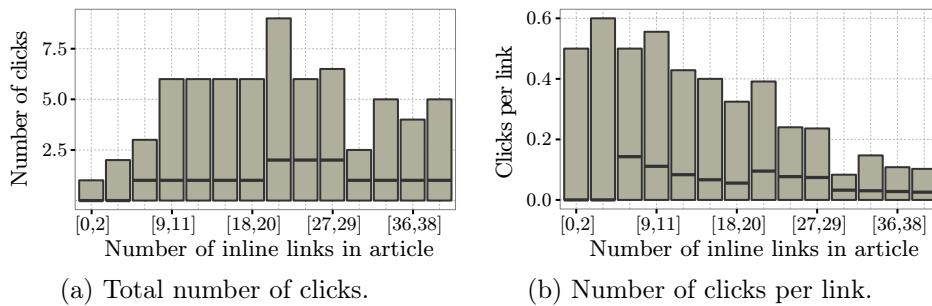


Figure 10.6: Number of clicks on inline links depending on the number of inline links of an article.

This suggests that (i) having less than ≈ 10 inline links per article may be wasting an opportunity, since users may be enticed to click to access related content by offering more links to them; (ii) having between 10 and 29 inline links per article does not result in more clicks, but simply spreads the clicks more; finally (iii) having more than 29 inline links may actually harm the user experience and make users less likely to click. This aligns with a user study reported in [10] showing that too many inline links can have detrimental effect on users' reading experience.

10.7.2 Types of Inline Links

Table 10.4 shows the popularity and performance depending on the types of inline links, using the predefined site categories in Section 10.3. We also report the percentage of articles containing a link of that type (“%Art.”).

Internal links appear in 87.73% of the articles that have inline links. These links include internal links to articles, and internal links to non-article pages (the latter includes links to topic pages, profiles of politicians or celebrities, *etc.*). Both categories of internal links occur in $\approx 60\%$ of the articles, but the popularity of internal links to non-article pages is higher. On the other hand, internal links to article pages have a higher performance than those that point to non-article pages (+80.39%, and -11.70% , respectively). However, we know from Section 10.6 that these links are not as frequently used as links on the front page of the news provider. We hypothesise that the provided links are not well presented or they do not cover the full information need of the users; in other words, their potential in driving users to consume more content has yet to be fully exploited.

Table 10.4: Popularity and performance of different types of inline links.

Link type	%Art.	Popularity	Performance
Internal	87.73%	72.20%	+13.48%
Article	59.76%	28.97%	+80.39%
Non-Art.	62.94%	43.22%	-11.70%
External	51.14%	27.80%	-14.58%
News Article	14.74%	3.68%	-29.66%
News Non-Art.	2.24%	0.46%	-15.19%
News Non-Top	27.33%	11.11%	-2.60%
Other	17.33%	4.11%	-9.34%
Organisation	11.61%	3.09%	-21.88%
Social Media	7.50%	4.15%	-90.89%
Multimedia	3.58%	0.75%	+60.75%
Reference	1.59%	0.45%	-46.53%

External links appear in 51.14% of the articles that have inline links. The most common type of external links are links to news sites outside our sample of top English news sites (“External News Non-Top”), but also links to popular news sites appear frequently. However, the link performance is -2.60% for links to less popular news sites, and -29.66% for links to popular news sites. This suggests that users are more attracted to less known providers, since they provide new information related to the story.

Links of types “External Other” and “External Organisation” also appear frequently in articles. For instance, the website of the Royal Astronomical Society of Canada¹³ was linked from articles related to the story about the “Draconid meteor shower”. Other articles link to valuable background information about a story; research studies related to the story “Volcanoes on Mars”¹⁴ or insurance information for “Obamacare”¹⁵. The performance of inline links to external other content is the third highest with -9.34% .

Links to reference websites (such as Wikipedia) and social media sites are less frequent and less likely to be used, particularly in the case of social media links. For both categories of links, the link performance is the lowest (-90.89% and -46.53% , respectively). Only for links to multimedia content we observe that the link performance is above 0. This suggests that users are interested in gaining more information from multimedia channels (such as YouTube). None of the classes of external links have a performance that compare to the internal links to articles pages.

¹³<http://www.rasc.ca>

¹⁴redplanet.asu.edu/?p=2389

¹⁵kff.org/health-reform/perspective/how-buying-insurance-will-change-under-obamacare

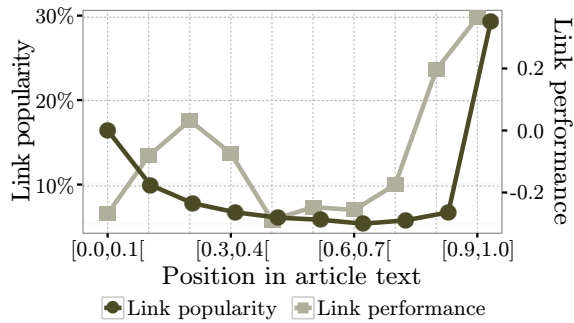


Figure 10.7: Popularity and performance of inline links depending on their position in the text.

10.7.3 Position of Inline Links

We examine the effect of the link position in relation to story-focused reading. We define the position of an inline link in the article text by counting the number of words occurring before the link. We then normalise the position between 0 (beginning of the text) and 1 (end of the text). Figure 10.7 depicts the popularity (left y-axis), and the performance (right y-axis) of inline links depending on their position.

We see that 30% of the links occur at the end whereas 16% of them appear at the beginning of the text. The remaining 46% are distributed within the article text. However, the performance of links located at the beginning of the text is very low (-28%), whereas the best performance is achieved with links at the end of the article text ($+35\%$). We hypothesise that users interested in a story (thus focusing on it), like to read the whole article first, before eventually deciding to read more articles on the same story.

To support story-focused reading, a good strategy seems to have a paragraph with inline links to related articles at the end of the article text, which is actually often the case for many articles. We also observed the same when restricted to inline links of types “Internal News Article”, “External News Article”, or “External News Non-Top”.

Interestingly, the inline links located between the upper 20% and 40% of the article text perform also well. A manual inspection of the data shows that these are links to multimedia content. Many news providers have articles embedding a picture with a link to a gallery in their upper part. However, we

could also find examples of solely text-based links that refer to multimedia content related to the article story.

Overall, we showed that the performance of inline links, which allow users to engage in focused reading (as they bring users to content related to the story of an article), can be affected by their type, their position, and how many of them are present on an article.

10.8 Effect on User Engagement

We know from Section 10.6 that users exhibit a different reading behaviour when focusing on a story; the reading sessions are longer and involve a larger number of news providers. We now investigate whether inline links, which can be viewed to promote story-focused reading, have an effect on user reading behaviour on a news site and hence the engagement with that site.

For a given news provider, we analyse the user activity on that site. In this context, a *provider session* is the collection of page views of a news reading session that belong to the provider site under consideration (views on article and non-article pages). A provider session is “story-focused” if the user clicked on at least one inline link (read at least two articles related to a same story). If the inline link brings the user to an external page (outside the provider), we refer to that provider session as “external-focused”.

Inline links and provider sessions. We first look at the relationships between inline links and provider sessions. The percentage of inline links in articles of a news provider correlates moderately with the percentage of story-focused provider sessions ($\rho = 0.62$, p-value<0.01). In addition, the percentage of external inline links correlates moderately with the percentage of external-focused provider sessions ($\rho = 0.56$, p-value<0.01). Therefore, providing inline links can lead to more story-focused reading within a news site. We now investigate whether this leads to higher user engagement with the news site as well.

Dataset. We focus on 50 news sites having users with at least one story-focused and one non-story-focused provider sessions, resulting in a sample of 57K users and 1M provider sessions. Restricting to users with both types of news reading ensure that any observed difference is not an artefact of user

browsing activity (*e.g.* users performing story-focused reading are always more engaged than users that do not focus on stories).

Engagement metrics. We measure user engagement with three metrics. For each news provider site, we calculate the average number of *page views* and the *dwelt time* per provider session of a user. We also calculate the loyalty of a user to that site, using *absence time*, which is the time elapsed between two provider sessions of a user. This metric was experimented in [72], where it was assumed that engaged users return sooner to a site, and hence their absence times are shorter. Here, we study whether story-focused reading has an effect on this metric (*e.g.* leads to shorter absence time). We calculate the percentage of provider sessions with an absence time below 12 hours, which represents users who come back to the same news site within that time.¹⁶

Figure 10.8 depicts the average dwell time, and the probability that the absence time is below 12 hours per provider, depending on the type of provider session.¹⁷ The x-axis represent the providers, ordered by increasing dwell time for non-story-focused provider sessions (represented by the line). The two types of dots represent the dwell time and absence time, respectively, for story-focused provider sessions (circle) and external-focused provider sessions (cross). A similar plot is obtained for the average page views metric, and therefore not shown here. Now, we discuss the results for these two types of sessions.

Story-focused provider sessions. The dwell time is higher for story-focused provider sessions, for almost all considered news providers. Only 3 (out of 50) providers have their corresponding average dwell time lower for the story-focused provider sessions. The average increase in dwell time from non-story-focused to story-focused provider sessions is 50%.

The same can be observed with respect to the loyalty metric. For 78% of the providers, we find that there are more users that return earlier after they have a story-focused provider session. The probability that users come back to the same news provider within the following 12 hours increases by 68%. The K-S test confirms that the differences are statistically significant ($p\text{-value} \ll 0.01$).

¹⁶The same results could be observed using 6, 24 and 36 hours.

¹⁷The axis values are removed for the sake of confidentiality.

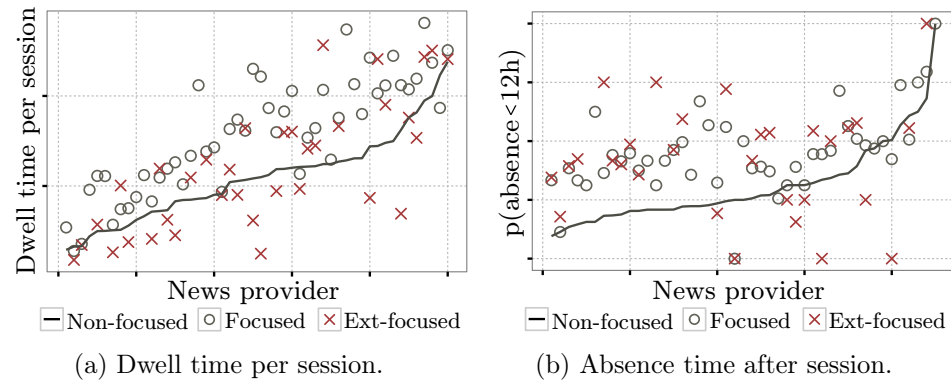


Figure 10.8: Session activity (dwell time) and loyalty (absence time) of users depending on the type of provider session.

External-focused provider sessions. Some providers do not offer enough inline links to external content, and we were not able to identify external-focused provider sessions for them (for these providers there are no values in the two plots in Figure 10.8).¹⁸ We focus on the remaining 35 news provider sites, consisting of 37K sessions and 31K users. We do not observe an effect on the dwell time (neither positive nor negative). The average increase is only 5.5%, and based on the K-S test we cannot confirm that the distributions are different (p-value=0.36).

Interestingly, for 70% of these news sites, the probability that users return within the following 12 hours increases (the average increase is 76%). The difference is statistically significant (p-value \ll 0.01). This suggests that the belief that links to external sites may hurt user engagement (with a site providing such links) is not founded. In many cases, users navigating to external sites when engaged in story-focused reading are more likely to return sooner to the news site. Their experience was positive, and such users are inclined to return to these sites sooner to consume more content.

Overall, providing links (inline) thus to promote story-focused reading has a positive effect on user engagement, in terms of time spent on the site, the number of articles read on the site, and loyalty to the site.

¹⁸These providers have on average only 3.5% external inline links, compared to the remaining 35 providers that have on average 6.6% external inline links.

10.9 Discussion

This research provides new insights about how users consume news online. We performed a large-scale data analysis that focused on a specific aspect of online news consumption: when users focus on a story while reading news, *i.e.* they read more than one article related to a specific story. We referred to this as *story-focused reading*.

Most studies investigating online news consumption have been concerned with how users read news on a specific news site, with the aim to find strategies to keep users engaged, *i.e.* reading many articles in each session. We add to this body of work a new dimension: We studied user reading behaviour across news sites when focusing on a story. We also showed that supporting story-focused reading is a good strategy to keep users engaged.

First, we showed that story-focused reading exists and that it is not a trivial phenomenon. Users decide to focus on a story, because they are interested in it. The popularity of a story, from both the providers' side and the readers' size, plays a minor role, as well as the amount of information provided per article measured by article length. We saw that users spend more time per article when focusing on a specific story, indicating that promoting story-focused reading might keep users engaged with a news site. We also showed that inter-site engagement increases when users focus on a story and the more articles they read about a story. The extent of inter-site engagement also depends on how users search for story-related articles, whereby search and social media sites are leading to a higher traffic flow between news providers than front pages or other news articles.

We could see that story-focused reading does not seem to be well supported at the article level, because in many cases the users utilise other means (*e.g.* front or section pages) to access further articles related to the story they are interested in. News providers should provide precisely for those users the means to engage in story-focused reading. By viewing an article, the users are already revealing that they are interested in the story of that article.

Motivated by this, we analysed how news sites promote story-focused reading, by looking at how they link their articles to other related content either published by them or other sources. We saw that 75.45% of the articles have inline links, which connect to other pages with related content (news articles and other types of pages). We found that users tend to click on links that bring them to other news articles within the same news site, or to articles

published by less known providers, probably because they provide new or less mainstream information. However, it is not a good strategy to offer too many such links, since this is likely to confuse or annoy users. We also found that users tend to click on links that are close to the end of the article text.

Finally, we showed that having links within the article text promotes story-focused reading and as a result keeps users engaged. The reading sessions of news sites that provide such links are longer. Additionally, users come back to these news sites sooner (shorter absence time). We also showed that promoting inter-site engagement by linking to external content does not have a negative effect on the engagement with the site; the reading time on the news site remains the same, and the absence time even decreases. As already observed in the Chapter 5 and Chapter 9, leaving a site does not necessarily entail less engagement, since users often return later on, and providing them with interesting or high quality content of other sites might even increase the engagement with the site. It should be, however, emphasised that this does not mean that news providers should just provide such links, but the right ones in terms of quantity and quality. As already discussed, the type, position, and amount of links play an important role.

Limitations. We discuss some limitations. We performed a non-overlapping clustering to define our stories. However, there are many stories that are related to each other (*e.g.* *Obamacare* and the threat of government shutdown), and this could be taken into account when studying story-focused reading.

Second, we did not consider all types of related links on an article page. Many news sites provide links in “related content” frames which are usually on the right side of the article text. Since we observed that users prefer to use inline links at the end of the article text, we assume that such frames are also performing well. They do not interrupt users while reading the article and they are usually clearly visible on the page.

Finally, we did not take into account how the quality of the related content influences user engagement with respect to the news provider. We expect that linking to low quality content will have a negative effect on users; they will not click on related content again. Also, it might influence the engagement of users negatively.



Discovering Story-related Content in Twitter

In the previous chapter we have shown that offering related articles and other information to news stories is a good strategy to keep users engaged with the news site. Due to this observation, we are now developing an application that can help news providers, their journalists and editors, in discovering story-related content and their curators in Twitter.

11.1 Introduction

Nowadays, users engage with a news provider not only through visiting the corresponding news site, but also through sharing its articles on social media platforms. This aspect of inter-site engagement makes social media platforms of particular interest to online news providers. For instance, news providers use Twitter and Facebook to spread news recently published on their websites, to assess the popularity of such news in different segments of their audience, but also to enrich the stories they publish on their websites. This is further accentuated by the seamless integration of social media platforms such as Twitter and Facebook into news websites, allowing easy content sharing and distribution.

A recent survey on 613 journalists over 16 countries revealed that 54% of them used online social media platforms (Twitter, Facebook and others) and 44% used blogs they already knew to elicit new story angles or verify stories they work on [189]. Indeed, Twitter is used by journalists and news editors of mainstream media sites to enrich their articles [66; 243]. They do

so by analyzing responses (post-read actions) to news articles [241; 253] as these can provide additional information about the news story, contained in discussions and opinions [134; 241], but also in URLs of related news published by other news sites [40].

Social media can be a powerful tool for journalists at multiple stages of the news production process: detection of newsworthy events, interpretation of them as meaningful developments, and investigation of their factual veracity [77]. Although Twitter users tweet mainly about daily activities, they also share URLs related to news stories [113; 285]. Indeed, 59% of them tweet or retweet news headlines [37], which account for 85% of the trending topics on Twitter [139].

In this work, we propose a radically new approach: We are leveraging *transient news crowds*, which are loosely-coupled groups that appear on Twitter around a particular news item, and where transient here reflects the fleeting nature of news. We then follow these crowds over time and show that parts of the crowds come together again around new newsworthy events related to the story. As an application of this observation, we design a method for automatically detecting content posted by them that is *related* to the original news story. The advantages are manifold: (1) the crowd is created automatically and available immediately, (2) we can account for the fleeting nature of news, (3) there is no need to maintain a list of experts or curators on Twitter.

Many users in Twitter also devote a substantial amount of time and effort to *news curation*. Digital news curation is an emerging trend where users carefully select and filter news stories that are highly relevant for a specific audience.¹ News curators can reach a high level of engagement and specialisation, becoming a sort of *distant witnesses* [34] of news stories of global significance. Twitter users can benefit from news curation by using the “lists” feature, that allows them to organise the people they follow into arbitrary topics; these lists are routinely used by many organisations including media companies to collect content around developing news stories. However, users still have to create such lists by hand and update them when deemed necessary.

We noticed that news crowds are made of users playing different roles. For instance, as in many online communities, a minority is actively engaged

¹<http://www.pbs.org/idealab/2010/04/our-friends-become-curators-of-twitter-based-news092.html>

(posting many articles), while the majority remains largely silent [173]. In the second part of our work, we therefore define and characterise some of the roles users play in news crowds. We then show that to some extent we can statistically model these roles. Our aim is to identify *news story curator*, that is, users who can be suitable curators for the story of a news crowd.

Our goals are therefore four-fold: (1) define the notion of “transient news crowds” on Twitter, (2) study their characteristics, (3) investigate how these can be exploited to discover story-related content posted on Twitter, and (4) define and model news story curators.

Transient news crowds and their news story curators can be used from journalists to enrich their articles by linking to relevant content (*e.g.* news articles, reports, videos) posted by them. This has been shown to have a positive effect on users’ engagement with the news provider (see Chapter 10). In addition, journalists can cover story *beats* incorporating the shifts of interest of the audiences that follow those beats. This represents an important step: given that journalists can be seen as members of an interpretive community [282] who come together to make sense of events and translate their importance, transient news crowds might represent individual news users demanding to be part of that interpretive community.

Even a casual examination of the data can show the potential of news crowds. For instance, on January 6, 2013, an article with title “Delhi police dispute India gang-rape account” was posted in the Al Jazeera English website and attracted 147 users who tweeted its link in the first six hours after its publication. Two days later, 27 of those users (18%) tweeted a link to a Huffington Post article with title “Father of Indian gang-rape victim speaks out”. If we were able to detect such articles automatically, we could generate a timely alert for the author of the first article pointing to the related article found by the crowd. Of course, we do not assume that *every* subsequent posted article will be related. Instead, we show that such related articles exist and that it is possible to detect them automatically.

The remainder of this chapter is organised as follows. Section 11.2 outlines related works. The datasets and its processing are described in Section 11.3. Section 11.4 defines and characterise transient news crowds. The models to identify articles and curators in a news crowd that are related to a news story are described and evaluated in Sections 11.5 and 11.6, respectively. Section 11.7 presents our discussion.

11.2 Related Work

Recommender systems. Twitter has been used as a source of news recommendations, typically by exploiting Twitter-specific features extracted from post-read social responses [4; 59; 207], tweets content (hashtags, topics, entities), users followees and followers, public timelines and retweeting behaviour. However these works aim at building personalised recommender systems, suggesting news articles based on the inferred topical interests of a user.

Our objective is entirely different, as we want to follow specific stories over time and offer related news articles to the authors of such stories. We want to provide journalists and editors a tool to discover new content that can complement or extend the one that they have produced.

Community detection. Many studies aiming at detecting Twitter communities around topics exist [92; 181]. The methods used rely on the extraction and incorporation of numerous features, such as user tweets [93; 284], but also user profiles and link similarity: how often two users retweeted, mention or reply to a common third person tweets [93]. The similarity of the tweet text, URLs, and hashtags have also been considered in the creation of such communities [284], as well as user mentions (addressing/replying), retweets, follower networks, and user lists [181].

Topic engagement (*e.g.* whether a user will join a discussion) has also been predicted [212; 270]. The content of tweets has been found to be a significant feature for this task, and retweeting the tweets of a user has been found to be a stronger indicator of topical interest than following a user.

Our approach is a natural complement to these works, which carefully craft a topically-focused community around a topic, and then assume that all the content produced by that community is on-topic. Instead, we put together a set of users that have a priori only one element in common (they tweeted a URL), and carefully filter the tweets they produce in order to find relevant on-topic content. Of course, both approaches can be combined.

User lists on Twitter have been used to detect communities [92]. Recent studies are concerned with recommending new tweets to a list [71], understanding the nature of curators, *e.g.* member and subscriber [87], and investigating users interests [127].

Our work can be viewed as a mean to automatically build such lists, which we recall are built manually, but accounting for the fleeting and volatile nature of news and with the aim to discover and recommend related news around a news story.

Expert finding. The availability of text collections in which authorship can be identified has generated a significant amount of activity around the topic of expert finding. The first approaches to expert finding were either text-based [52] or network-based [1]. Both paradigms have evolved over the years, and recent approaches combined them (*e.g.* [64]). Expert detection in Twitter has become an active research topic [133], especially with the increasing usage of Twitter as a service for news aggregation and consumption. Twitter experts, called curators, collect high valuable and informative news and other content around a topic. They also are known to identify interesting news (that end up becoming popular) early on [102].

The detection of curators is a difficult challenge mainly caused by the dynamic nature of Twitter. For instance, Pal et al. [203] argued that network-based features (*e.g.* follower-network) are not always suitable, because the lifetime of a topic can be very short. In addition, users are followed for other reasons than their topical expertise, thus reducing the effectiveness of network-based features to detect experts in Twitter [271]. Network features based on the retweet network in combination with content features were shown to better reflect the dynamic nature of Twitter and as such more suitable for the task of detecting experts [133].

Contextual data (as contained in the user profile including the user name) have to be used carefully, since user studies showed that the name of the user may bias the judgment. Indeed, Pal et al. [203] demonstrated that users rank topic-related tweets from celebrities as more interesting and authoritative than when the same information is tweeted by non-celebrity users. On the other hand, contextual data provides useful information such as the lists a user belongs to [90; 162; 259].

User lists are a widely used Twitter feature that allows users to group other users, for instance, around a same topic. Wagner et al. [259] demonstrated that features based on the user lists perform best, compared to content-based features based on recent tweet/retweets and features based on the profile. Approaches to automatically extend or create these lists exist as well [29]. Nowadays, these lists are also used to filter valuable content for journalists. For instance, *Storyful* is a news agency that provides user lists, developed by journalists for journalists.

Table 11.1: Dataset characteristics: number of articles, users, and tweets.

Dataset	Articles	Users		Tweets	
		Total	Per crowd	Total	Per crowd
BBC World Service	75	13.3K	177	35.5K	201
BBC News UK	141	13.1K	92	47.8K	339
Al Jazeera English	155	8.3K	53	24.0K	154

Our approach is different because our aim is not to develop a new expert detection approach for Twitter. We took the perspective of journalists and editors who are interested in understanding the users who are tweeting their articles, and want to detect those users that could provide further content to the story of the news article. We refer to these users as news story curators.

11.3 Datasets

We describe the data used in our work, how it was created, and various processing performed.

Data extraction. We collected news articles published in early 2013 on two major online news websites, BBC and Al Jazeera English (AJE). The news websites represent large media organisations, seeking adoption of their content in a wide range of international markets. From the BBC, we collected separately articles from World Service (BBC-WORLD) and BBC UK (BBC-UK), each forming a different dataset. We downloaded periodically the homepage of each website, from which we sampled at random a subset of news articles. We focused on the headline news: opinions, magazine and sport news were not included. The sampled articles cover a variety of stories such as Obama’s inauguration, the conflict in Mali, the pollution in Beijing, and road accidents in the UK.

For each of the sampled articles, we started a process that used Twitter’s API² to periodically find tweets including that article’s URL. The earliest tweets followed almost immediately the publication of the article, since each of these news organisations disseminate their content via their own twitter account(s) (*e.g.* @BBCWorld, @AJEnglish). We define the crowd of a news article as the set of users that tweeted the article within the first 6 hours after the first tweet on that article. We selected this time period because it encompasses about 90% of the tweets an article receives (87% for BBC-

²<http://dev.twitter.com/>

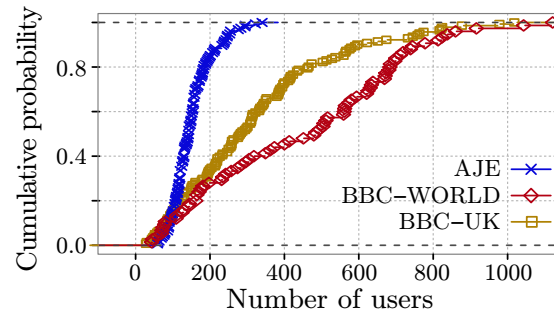


Figure 11.1: Distributions of the number of users per crowd.

WORLD, 91% for BBC-UK, and 91% for AJE). We followed users on each crowd during one week, recording every public tweet they posted during this period.

Data filtering. In Twitter there is a substantial amount of spam. Spammers routinely abuse Twitter to promote products and services. Successful spammers attract followers and bypass filtering mechanisms by posting a mixture of reputable tweets and advertising [22]. Spam can negatively affect our results, and given that the Twitter API has strict rate limitations, it can also reduce the coverage of our dataset by forcing us to waste our quota downloading useless tweets. Hence, it is important to define some criteria to filter out at least the most active spammers.

Informed by previous works [22; 261], as an heuristic to remove spammers, we removed users with an abnormally high tweeting activity (98 tweets per day), whereby most of the tweets were retweets (90% retweets) or tweets containing URLs (92% URL-tweets). We also examined manually the most prolific accounts and defined a blacklist of high-throughput automatic accounts that do not focus on any particular region, topic, or news provider. We removed only accounts having clearly anomalous behaviour, and tried to keep our manual intervention to a minimum, discarding less than 5% of the users in total.

Finally, news articles with very small crowds (lower 15% of the distribution) or very large ones (upper 5% of the distribution) were excluded. We kept articles with 50–150 users for BBC-WORLD and BBC-UK news articles and 70–360 users for AJE. The resulting number of articles, the sizes of the crowds, and the number of tweets collected for each dataset are summarised in Table 11.1. As shown in Figure 11.1, these distributions are very skewed

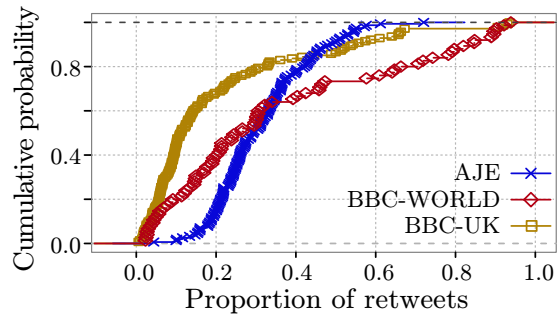


Figure 11.2: Proportion of retweets during each crowd’s creation.

and there are crowds that are substantially larger than the average, as well as users that are substantially more prolific than the average. We observe that the crowds around articles in AJE are smaller than the ones of BBC-WORLD and BBC-UK, a reflection of the different sizes of their audiences.

Shortened URL handling. The limitation of number of characters in tweets is viewed as one of the key elements of the success of Twitter as a sharing platform. However, it also imposes constraints for users who want to post URLs, which are usually long strings. Hence, a number of *URL shortening* services have appeared in recent years, providing on-demand URL alias such as “<http://tinyurl.com/2g774x>”. URL shortening services typically generate different shortened URLs for different users, given the same input URL. Expanding shortened URLs requires at least one network access, thus creating a bottleneck for many applications that should be avoided when possible.

We therefore expanded only a subset of the URLs appearing in our dataset. To determine this subset we rely on the text of the tweets containing the URL. That text is stemmed, stopwords are removed, and word bigrams are extracted; the latter are used as the tweet representation. Two URLs are considered equal if they appear in tweets having a Jaccard similarity of at least θ under this representation. The threshold is determined by using the outcome of a crowdsourcing task in which 300 random pairs of tweets from our collection were annotated by humans (details of the crowdsourcing service used are given in Section 11.5.3). We set $\theta = 0.25$, which has a precision and recall of 0.84 on this test set.

A shortened URL, without the need to be expanded, is thus represented as a cluster (computed by a greedy clustering algorithm) of equal URLs as cal-

culated above. Only one of the URLs in each cluster needs to be expanded, and the frequency of a URL is the number of tweets in its cluster. This definition of frequency is used in the remainder of this chapter, particularly in Section 11.5.4 in the task of discovering popular related content.

11.4 Crowd Characteristics

To the best of our knowledge this type of transient crowd in the news domain has not been studied in depth. We summarise key observations about the characteristics of these crowds in terms of their creation, members and dynamics.

11.4.1 Crowd Creation

There are two main mechanisms by which a user can tweet about a news article and hence become a member of a crowd: direct tweets and re-tweets. *Direct tweets* can be done by the user by clicking on a “tweet” button provided by the news website, or by using a bookmarklet, or by copying and pasting the URL in a Twitter client. *Re-tweets* are created by users in a Twitter client or directly on the Twitter website, and correspond to re-posting what other users have shared.

Figure 11.2 depicts the proportion of retweets for our three datasets. This proportion is basically below 0.4. This indicates that a large proportion of the activity around a news URL on Twitter can be traced back directly to the news website, and not to word-of-mouth/propagation effects in Twitter. However, in AJE we observe a stronger word-of-mouth effect than in the other two sites, which is consistent with previous observations [169].

11.4.2 Crowd Members

We look at the behaviour of the users belonging to news crowds during the one-week period following their creation (starting with the first tweet about a news article). In Figure 11.3a we plot the distribution of the average number of tweets per day of crowd members, which peaks at around 40 tweets/day for AJE and 60 tweets/day for BBC-WORLD and BBC-UK. In any case, these are unusually high numbers, given that the overall average is around 2.5 tweets/day.³

³By the end of 2012, CNET reported that the number of tweets per day was close to 500 million (<http://cnet.co/U3h0UW>), while the number of active users, according to The Guardian, was around 200 million (<http://gu.com/p/3cjvf/tw>).

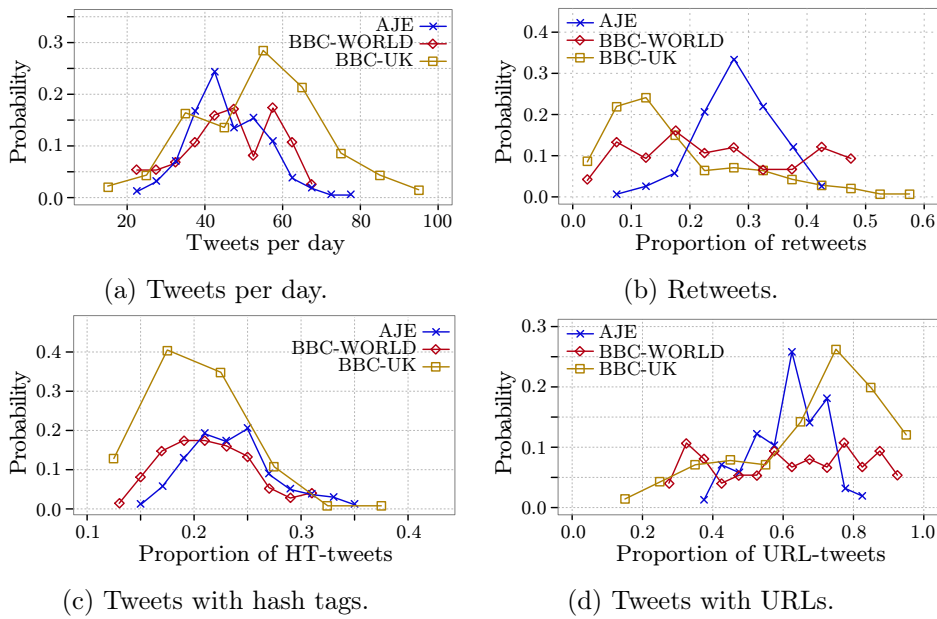


Figure 11.3: Distributions of number of tweets per day and different type of tweets.

Indeed, after our spam filtering heuristics (Section 11.3), crowds still include many Twitter accounts that post tweets automatically but are not spammers. These include corporate accounts operated by the news networks themselves, such as `@BBCWorld` and `@AJELive` (from Al Jazeera). They also include services provided by third parties, such as `@bbcnews_ticker` that tweets all the news in the BBC news ticker, `@AmmanHashtags` that automatically re-tweets news mentioning the capital of Jordan, and `@TwittyAlgeria` that tweets news containing the word “Algeria” extracted from a variety of sources.

At the same time, there are several accounts that do not seem to be operated automatically and that are examples of good news curators. For instance, `@thomas_wiegold` has a few thousand followers and manually curates a set of conflict stories around the globe and adds commentary aimed at a German-speaking audience.

Crowds can also be described by the characteristics of the tweets posted by their members. The fraction of tweets that are re-tweets, shown in Figure 11.3b, is consistent with Figure 11.2, showing a large proportion of re-

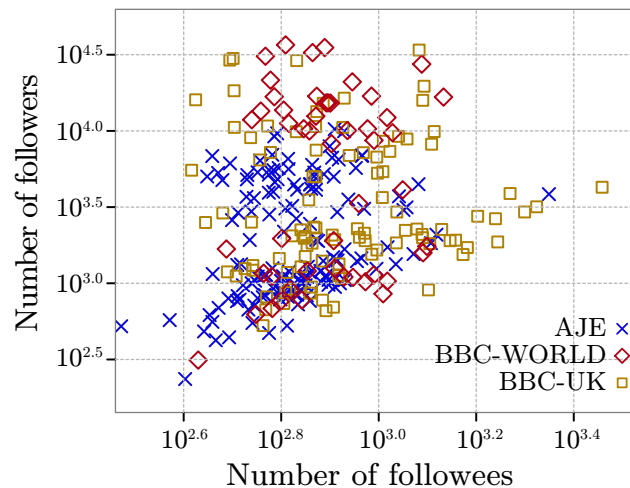


Figure 11.4: Average number of followers and followees of users per crowd. Each data point corresponds to one crowd.

tweets in AJE. The fraction of tweets containing hashtags (Figure 11.3c), or URLs (Figure 11.3d) indicates that in comparison with the other datasets, users of BBC-UK use slightly less hashtags (peak at 0.2 vs. peak at 0.25) and have a higher proportion of tweets with URLs (peak at 0.8 vs. peak at 0.6).

Figure 11.4 depicts each crowd from the perspective of the average number of followers and followees of its members. We observe that crowds in BBC-WORLD and BBC-UK have on average a larger number of followers than those in AJE. Overall, these values are relatively high considering that a majority of Twitter users have less than 50 followers.⁴

The average is dominated by crowd members having an extremely large number of followers, such as some of the accounts we have mentioned. For instance, `@TwitzAlgeria`, despite being evidently generated automatically, has over 240,000 followers (as of March 2013). Even if some of these followers were in turn automatically-created accounts, these large numbers indicate that users perceive their tweets as valuable, because otherwise they would have ceased to follow them (“unfollowing” an account in Twitter is as easy as following it). In other words, automatically generating/aggregating content does not seem to be perceived a priori as negative by Twitter users. Therefore, we do not attempt to remove automatic users from our crowds, but we do weight their influence carefully (as we show in Section 11.5.4).

⁴<http://www.beevolve.com/twitter-statistics/>

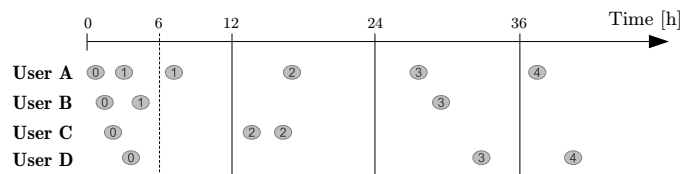


Figure 11.5: Depiction of our assignment of slices to tweets in the data. Each row corresponds to a user and each oval to a tweet, numbered with the time slice it belongs to. All the tweets containing the original URL are assigned to slice t_0 , and must be posted within 6 hours to grant crowd membership to its user. Subsequent tweets are assigned to other slices as per the diagram.

Recurring crowd members. Crowd members on a given website often overlap. About 74% (sd=0.13) of users who tweet an article in AJE tweet at least one other article in AJE during our observation period. Something similar happens with BBC-WORLD and BBC-UK, where respectively 61% (sd=0.24) and 75% (sd=0.22) of crowd members tweet more than one article. Again, these recurring crowd members include a mixture of automatic accounts but also manually-operated ones that choose to tweet from the same sources repeatedly. This reinforces the need to weight their influence carefully in the news discovery task.

11.4.3 Crowd Dynamics

To study the dynamics of crowds over time we discretise time into slices of a fixed size. We illustrate this in Figure 11.5 for an example set of four users. The tweets that create a crowd are assigned to the slice t_0 and are posted within 6 hours of each other. The remaining tweets from these users are assigned to a time slice according to the time passed since that first tweet. We perform this assignment independently in each of the crowds of our three datasets.

Time granularity. The choice of the appropriate time granularity for time slices depends on the application. In our case, we are interested in the news discovery problem described in Section 11.5, and hence, this problem informs our choice of a time granularity. We focus on the phenomenon of *topic drift*, by virtue of which each crowd “disperses” in terms of the content of their tweets. We can quantify this dispersion by first measuring the expected similarity between tweets in a time slice, and then observing

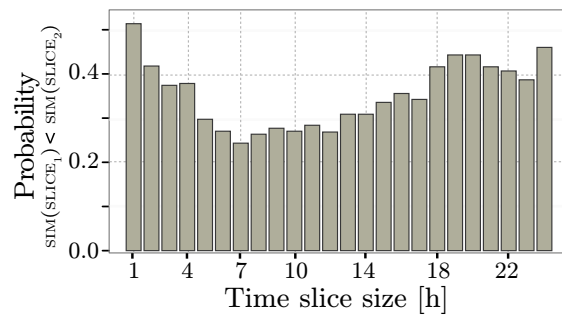


Figure 11.6: Time granularity: Probability that a crowd’s tweets become more similar on the second time slice (compared to the first time slice) for different choices of time granularity.

if this expected similarity changes over time. The similarity of two tweets is measured using the Jaccard coefficient of their representations as bags of words after stemming and stopword removal (see Section 11.3).

Over time, we expect that the average similarity becomes smaller. In particular, we expect that given our choice of time granularity, tweets on the first time slice of a crowd are more similar to each other than tweets on the second time slice of the same crowd. With this in mind, we study different time granularities ranging from 1 hour to 24 hours, and measure the probability that in a crowd the average similarity on the second slice is (contrary to our expectations) higher than the average similarity on the first slice.

Figure 11.6 shows the results for this test. For small granularities (*e.g.* 1 hour) the probability is close to 0.5, indicating that using that time granularity crowds can get either farther apart or closer together. This can be viewed as random, and any values above or below can be considered as a signal.

For granularities between 7 and 12 hours a minimum of less than 0.3 is attained, indicating that crowds have at least a 70% chance of becoming more dispersed in the slice t_2 with respect to slice t_1 . We chose a time granularity of 12 hours in the remainder of the chapter, since it is easy to interpret. In the Figure we observe that for larger time granularities, we return slowly to random behaviour, reaching 0.5 at granularities of 18-24 hours.

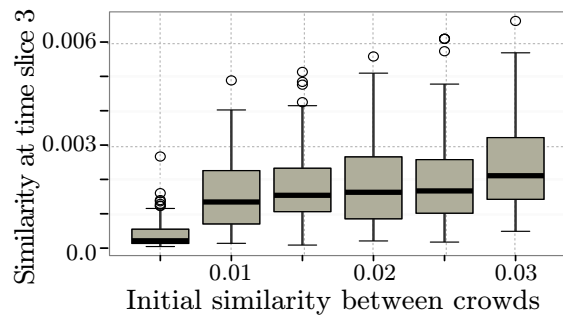


Figure 11.7: Correlation test: Distribution of similarity at slice t_3 for pairs of crowds at different levels of similarity at slice t_0 .

Correlation test. Next we must determine if at least part of the activities of a crowd are related to the article that created each crowd. In order to do this, we conduct a randomised test. We consider random pairs of crowds whose first slice overlaps (*i.e.* the original articles are posted within 12 hours of each other). First, we look at the similarity of the original articles, by measuring the average similarity of tweets containing the original URL (belonging to slice t_0 in both crowds). Second, we perform the same measure in the slice t_3 of both crowds. This test attempts to answer the following question: if two articles posted today are similar to each other, will users who tweeted about those articles tweet about similar things tomorrow?

The correlation obtained in general between both similarities is $r^2 \approx 0.4$. Figure 11.7 depicts the distribution of similarities in slice t_3 for different buckets of similarity at slice t_0 . We can see a clear trend in which articles that are not similar to each other rarely have crowds that tweet about the same topics in the future, while this often happens in crowds originating from articles that are similar to each other. This clearly shows that crowds are not formed randomly. Next, we use them for a news discovery task.

11.5 Crowd-based News Discovery

This study is motivated by the intention of discovering content related to a news story with the help of news crowds. In this section, we describe a method for performing such discovery. We formulate the discovery task as follows: *given a news crowd and a time slice, find URLs in that time slice that are related to the article that created the crowd.*

A number of steps are followed. First, we extract from each slice the most frequently posted URLs, as described in Section 11.5.1. Next, we use a supervised learning approach, which we explain in Section 11.5.2. The training dataset is described in Section 11.5.3. Three types of features were employed, frequency-based, text-based and user-based, described in Section 11.5.4. Our results are discussed in Section 11.5.5, and we also suggest an application to crowd visualisation over time task in Section 11.5.6.

11.5.1 Candidate Generation

We recall that given a URL (article) around which a crowd has been formed, the aim is to discover articles (their URLs) related to the original article. The first step is to extract a pool of candidate URLs from all the URLs tweeted by the crowd. In each time slice, we generate the top URLs having the largest frequencies, where the URL frequencies are computed using the method described in Section 11.3. We remove all URLs having frequency less than 3. This still yields a substantial number of candidates, 41.2 (sd=23.8) per time slice on average for BBC-WORLD, 54.8 (sd=23.8) for BBC-UK, and 15.7 (sd=4.7) for AJE.

Many of these candidate URLs are not related to the original article. We illustrate this with an example using two articles published on AJE on January 13th, 2013. Both articles correspond to ongoing armed conflicts in the Middle East (“Syria allows UN to step up food aid”) and Africa (“French troops launch ground combat in Mali”). We identify the crowds of each story and follow them during 14 time slices of 12 hours each, *i.e.* one week. Next, we manually assess whether each candidate is related to the original story or not. The result of this manual process is shown in Table 11.2.

For the crowd of the story on Syria, we can see that the number of candidates that are related to the original story consistently exceeds the number of candidates that are not related. For instance, in time slice t_5 we have five related candidates (including stories about the Taftanaz Military Airport, the Kilis refugee camp, *etc.*) and one unrelated candidate about a hostage crisis in Algeria.

For the crowd of the story on Mali, there are actually more unrelated candidates than related ones. Still, some time slices such as t_4 have three related candidates (two about the movements of troops in Mali and one about a related statement by French President Hollande) and one unrelated candidate about the hostage crisis in Algeria.

Table 11.2: Example of candidates found for two stories published on January 13, 2013. A candidate is a URL posted by 3 users or more during each of the time slices ($t_1 \dots t_{14}$). We include the number of candidates related to the original story, and the number and topics of those that are not related.

	Syria allows UN to step up food aid			French troops launch ground combat in Mali		
	Rel.	Not rel.	Examples	Rel.	Not rel.	Examples
t_1	7	0		1	3	Zero Dark Thirty (film), Algeria×2
t_2	7	0		1	1	Spain
t_3	9	0		0	0	
t_4	5	1	Algeria	3	1	Algeria
t_5	5	1	Algeria	1	0	
t_6	5	2	Crabtree (football), Iran	0	2	Manti Te'o (football), Algeria
t_7	8	1	Algeria	1	1	Chardy (tennis)
t_8	9	4	Mali, Obama, Davos, Batman (film)	1	4	Algeria×2, Soccer, Israel
t_9	8	0		1	1	Algeria
t_{10}	13	2	Iraq, Federer (tennis)	0	1	Flanders
t_{11}	10	1	Obama	1	3	Algeria×2, MLK
t_{12}	10	0		0	1	Algeria
t_{13}	5	2	Lady Gaga (artist), Algeria	1	2	Djokovic (tennis), Jordan
t_{14}	13	2	Beyonce (artist), Palestine	1	0	
	114	16	Total	12	20	Total

With a less restrictive definition than adopted here, the news on the Algerian hostage crisis could be considered as related to the news on the French troops in Mali, because the main demand of the kidnappers was the end of the French military operations in Mali. In this case, we would retrieve a total of 22 related and 10 unrelated candidates. This shows that how we define “relatedness” has an effect on the results.

However, the proportion of related candidates is still larger for the story on Syria. There can be many reasons for the differences, one being that the conflict in Mali is more recent than the one in Syria, hence the latter has many more stories, and a more cohesive crowd of users following the news related to it. It is however clear that relying solely on frequency information (URLs in our case) will often not be sufficient to identify related stories. Other features are important and need to be incorporated, as described next, using a learning approach.

11.5.2 Learning Framework

We experimented with several learning schemes on our data and obtained the best results using a random forest classifier as implemented in Weka.⁵ Given the large class imbalance, we applied asymmetric misclassification costs. Specifically, false negatives (classifying a relevant article as non relevant) were considered five times more costly than false positives; values close to this number did not change substantially the obtained results. For consistency and simplicity, we use the same cost across the three datasets, even if their priors are different.

We use standard evaluation metrics including precision, recall, and AUC, all measured after ten-fold cross validation. Given that the targeted users of this system (journalists) do not expect nor need to have perfect results, we decide to aim for a level of precision close to two-thirds, since we considered it would be satisfactory for them to see twice as many related content than unrelated content. Hence, a key metric in our evaluation is the *recall at two-thirds precision*, which measures the probability that related content is found in our system, if we allow it to generate at most one-third of unrelated content in its output.

11.5.3 Training Data

We collected about 22,500 labels for about 7,500 training examples through Crowdfunder, a crowdsourcing provider that provides an interface to a variety of crowdsourcing platforms.⁶ We sampled uniformly at random 160 crowds: 80 from AJE, 40 from BBC-WORLD, and 40 from BBC-UK. For each crowd, we selected 5 slices at random, and up to 10 random candidates (URLs having a frequency of 3 or more) from each selected slice.

For each candidate, we showed to three different crowdsourcing workers a sample tweet from the original story and a sample tweet from the candidate URL, and asked them to annotate the pair as follows (see Appendix A.3.1 for detailed instructions):

- [Q1] Please indicate how these two stories are related: strongly, weakly, or not related.

⁵<http://www.cs.waikato.ac.nz/ml/weka/>

⁶<http://www.crowdfunder.com/>

We merged the first two classes as simply *related* for the purpose of our experiments. We ignored the pairs for which the confidence (based on the agreement of the workers) was less than 0.8 and the label was *not related*, as these were almost always borderline cases and are not useful for training or evaluation purposes. Overall, the assessors determined that for BBC-WORLD 4.9% of the candidates were related, for BBC-UK 8.2% and for AJE 9.3%.

The ratio of weakly related candidates per strongly related candidate varies greatly across datasets: 1.6:1 for BBC-WORLD, 8.5:1 for BBC-UK, and 0.9:1 for AJE. In other words, while in AJE the assessors found candidates that were strongly or weakly related in roughly similar proportions, in the case of BBC-UK there are more than eight weakly related candidates for each strongly related one. This in fact has an effect on the performance obtained for BBC-UK, as described in Section 11.5.5.

11.5.4 Features

We describe the three sets of features employed in our learning algorithm. We employ features, inspired from related works (see Section 11.2), and also from the observations reported in Section 11.4.

Frequency-based Features

For each candidate URL we compute its relative frequency, *i.e.* the frequency of its URL divided by the frequency of the most frequent URL in the slice (we name this feature `CandidateNormalisedFrequency`).

As we described in Section 11.5.1, even candidates having a high frequency are often not related to the original news item. Often breaking news about events of global significance appear in many crowds at the same time. To alleviate this problem, we incorporate a feature, analogously to the inverse document frequency in Information Retrieval, that measures how specific is a candidate with respect to a given crowd. If there are n crowds that have time slices that overlap with a candidate appearing in n_c of them, then `CandidateInverseDocFrequency` = $\log(n/n_c)$.

We also observe that repeatedly the top URLs on a given slice can be traced back to prolific users (such as those mentioned in Section 11.4.2) that post hundreds of tweets per day. These observations inform the design of the user-based features described in this section.

Text-based Features

To remove candidates not related to the original story, we employ a text-similarity approach. We use the same representation of a URL that we used to compute its frequency: a cluster of tweets that contain variants of a URL. Given this representation, we compute the similarity between two URLs by concatenating all the tweets in each cluster in a document, and compute the Jaccard similarity between such documents. Since this approach does not require the web page content of the original news article and the candidate URLs, we are able to access non-textual candidates such as videos, pictures or podcasts. Moreover our approach is computationally more efficient as we deal with less content.

First, we measure how similar are the tweets containing the candidate URL to the ones containing the article that created each crowd. We compute four features based on different representations of the content: word unigrams (`SimOriginalUnigrams`), word bigrams (`SimOriginalBigrams`), hash tags (`SimOriginalHashtags`) and capitalised terms (`SimOriginalCapitalised`). The latter is equal to word unigrams except that only words starting with a capital letter are considered – a heuristic that is effective in our dataset given the news headlines writing style.

Second, we measure how similar are the tweets containing the candidate URL to other tweets that appear in candidates from other crowds. We consider only the slices of the other crowds that overlap with the candidate's slice and use text similarity measures to compute how unique is a candidate with respect to a given crowd. Again, we computed four features based on unigrams, bigrams, hashtags and capitalised terms, but determined through experimentation that only one of them was significant: `SimOthersHashtags`. In total, we used 5 text-based features.

User-based Features

Based on the analysis of Section 11.4, in particular the presence of prolific automatic accounts, it was deemed important to consider features related to the users that contributed each candidate. We therefore incorporated weighted frequency features, in which each user that posts the URL of a candidate contributes a “vote” to that candidate that is weighted by a user-dependent feature. The purpose of these weights is to increase the influence of users that are focused in a single topic, and conversely reduce the influence of users who post tweets about many different topics. Additionally, we want

Table 11.3: Evaluation of the discovery of related articles, in terms of area under the ROC curve (AUC) and recall at 2/3 precisions (R@2/3). Each row corresponds to a set of features. Empty cells indicate that a set of features is unable to attain 2/3 precision.

Features	#	AJE		BBC-WORLD		BBC-UK	
		AUC	R@2/3	AUC	R@2/3	AUC	R@2/3
Freq	2	0.65	-	0.64	-	0.54	-
Text	5	0.87	0.40	0.85	0.44	0.66	-
User	11	0.81	0.30	0.70	-	0.64	-
Freq+Text	7	0.89	0.62	0.88	0.52	0.66	0.04
Freq+User	13	0.79	0.32	0.72	-	0.64	0.11
Text+User	16	0.92	0.66	0.80	0.43	0.73	0.14
All	18	0.92	0.72	0.85	0.49	0.71	0.14

to increase the influence of users who have attracted a significant number of followers.

Specifically, we consider that a user voted according to (i) its ratio of followers to followees, `WeightedSumFollowerFollowees`, (ii) the inverse of the number of crowds s/he belongs to, `WeightedSumInverseCrowds`, and (iii) the inverse of the number of distinct sections of the crowds s/he belongs to, `WeightedSumInverseSections`.

For the latter, *sections* correspond to different topics/regions in the news websites we work with, and we associate crowds to a section by looking at the prefix of the path of the article originating each crowd. For instance, articles under <http://www.bbc.co.uk/news/wales/> correspond to the section “Wales” of BBC-UK. In websites organised in a different manner, other ways of defining sections may be necessary.

Additionally, we characterise the activity of users contributing to a candidate by averaging the following quantities in each crowd: their overall volume of tweets per day (`UserTweetsDaily`), their number of followers and followees (`UserFollowers` and `UserFollowees`), and how many tweets they have favoured (`UserFavorites`).

We also obtained statistics from their tweets by computing the fraction of their tweets that contains a re-tweet mark “RT”, a URL, a user mention or a hashtag (respectively `UserFracRetweets`, `UserFracURL`, `UserFracMention`, and `UserFracHashtag`).

Table 11.4: Aggregated ranking of features by importance (most important first) across the three datasets. Our model uses frequency-based (F), text-based (T), and user-based (U) features.

1	[T]	SimOriginalBigrams	10	[U]	UserFavorites
2	[T]	SimOriginalCapitalised	11	[U]	WeightedSumFollowerFollowees
3	[U]	WeightedSumInverseCrowds	12	[F]	CandidateNormalisedFrequency
4	[T]	SimOriginalUnigrams	13	[U]	UserFracHashtag
5	[F]	CandidateInverseDocFrequency	14	[U]	UserFracMention
6	[U]	UserTweetsDaily	15	[U]	UserFracURL
7	[T]	SimOthersHashtags	16	[U]	UserFollowees
8	[U]	WeightedSumInverseSections	17	[U]	UserFracRetweets
9	[U]	UserFollowers	18	[T]	SimOriginalHashtags

11.5.5 Results

The performance of our automatic method for discovering related content is shown in Table 11.3. This method was applied over the three most frequent URLs on each slice. This was found to be much faster than considering all candidates and, in addition, it led to a similar accuracy than considering them all – this means that this set usually contains the related news that matter.

We include results with the 2 frequency-based features, the 5 text-based features, the 11 user-based features, and combinations of them. We observe that as expected the combination of these features yields the best performance. User-based features are valuable, even if they cause a decrease of 3 points of recall (at the desired level of precision) for BBC-WORLD; they bring a substantial increase of 10 points for AJE and BBC-UK.

In the case of BBC-UK we discover 14% of the related content using our method. In the cases of AJE and BBC-WORLD we can discover half or more of the related news in each crowd at the same level of precision. The difference in performance can be partially explained by the high proportion of weakly-related content in BBC-UK (see end of Section 11.5.2), *e.g.* road accidents that are related to other road accidents but often do not belong to long-standing issues such as the ones covered by BBC-WORLD and AJE.

Our features largely complement each other, as several feature selection methods failed to produce consistent gains in terms of these metrics. We can apply a feature selection method to BBC-WORLD to make it perform better, but if we use the same feature selection method in the other datasets we decrease the effectiveness of our models. In a real-world deployment of such a system, it will therefore be important to identify the particular

combination of features that lead to the best performance on a specific dataset.

We observe that across datasets some features are always valuable, while others contribute only in some cases. Table 11.4 shows the features sorted by decreasing order of importance, using an aggregation (Borda count) of their rankings by chi-squared tests in each dataset.

The most important features include the similarity to the original story, as well as measures of how unique is the association of the candidate URL and its contributing users to the specific story's crowd. This interesting result is well aligned with previous works (tweet content as an important feature) but also with the characteristics of the transient news crowds we reported in Section 11.4.

11.5.6 Application

The discovery of related content can help summarizing the evolution of a crowd over time, as we illustrate briefly in this section. We use as example the article “Central African rebels advance on capital”, posted in AJE on 28 December, 2012.

We considered a baseline that selected up to 3 candidates, posted by at least 3 users each, based on their frequency. This is the method employed to produce Table 11.2. We compared this against our method that classified each of these candidates as relevant or not. We took the output of both systems and used frequent words used in the tweets containing each URL to create word clouds for the time slices t_1 , t_8 and t_{14} of this crowd, as show in in Figure 11.8. As usual, font sizes are proportional to word frequencies.

The word clouds show that the candidates filtered by our method are related to the original story. Four days after the news article was published (t_8), several members of the crowd tweeted an article about the fact that the rebels were considering a coalition offer. Seven days after the news article was published (t_{14}), crowd members posted that rebels had stopped advancing towards Bangui, the capital of the Central African Republic. If we do not filter the candidates (using our method) we find articles on a wide range of topics that are popular, but weakly related or not related at all to the original news article. The method we use to discover related articles can yield a method for representing the evolution of a news story over time.

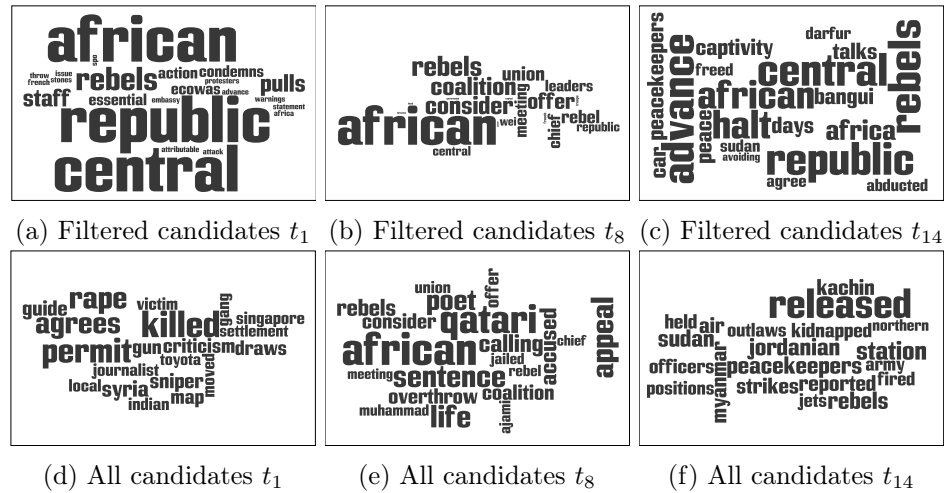


Figure 11.8: Word clouds generated for the crowd on the AJE story “Central African rebels advance on capital”, by considering the 20 most frequent terms appearing in stories filtered by our method (top) and on the top-3 candidates by frequency (bottom).

11.6 News Story Curators

In this section we define and characterise the roles of users in a transient news crowds. We then present an approach for automatically finding news story curators in a news crowd (*i.e.* for the corresponding story). The approach can be formulated by the following task: *Given a news article and its crowd of users, our goal is to identify which of those users can be suitable curators for the story the article belongs to.*

We proceed as follows: In Section 11.6.1 we define and characterise some of the roles users play in transient news crowds. The model and the training dataset are introduced in Sections 11.6.2 and 11.6.3, respectively. We employ 9 features reported in Section 11.6.4. Our results are discussed in Section 11.6.5, and we also present a precision-oriented evaluation of the model in Section 11.6.6.

11.6.1 Concepts

Digital curation is a broad field concerned with the management and preservation of digital data, specially considering future re-use [276]. We focus on the role of *online content curator*, which has been defined as “someone who

continually finds, groups, organises and shares the best and most relevant content on a specific issue online.” [24]

News story curators. We consider that among the users in a news crowd, some of them may be online content curators of the corresponding news story, that is, they follow the development of the story. Our aim is to identify such curators, which we call *news story curators*. A famous example for this type of news curator is Andy Carvin (@acarvin), who mostly collects news related to the Arabic world, and became famous for his curatorial work during the Arab Spring [34]. As a news story curator for the Arab Spring, he aggregated reports in real time and tweeted sometimes thousands of tweets per day.

A manual analysis of our data and the characteristics of Twitter curators in general revealed different types of curators and we assume that these types are important for our work. We present next the dimensions in which curators can be divided.

Topic-focused/unfocused. We observed two types of users that are intensely engaged with news content in social media. We call them focused curators and unfocused curators. A *topic-unfocused curator* is a user that collects contents about diverse topics, becoming a news provider of sorts, disseminating news articles about breaking news and top stories. For instance, @KeriJSmith, a self-defined “internet marketer” tweets about various interesting news on broad topics. A *topic-focused curator* is a more selective user, who collects interesting information with a specific focus. This focus is usually a geographic region or a topic. For instance, @chanadbh tweets about news related to Bahrain, whereas @brainpicker collects contents about art and design. Topic-focused curators play a pivotal role in the filtering and dissemination of news, and constitute a first line of defense against information overload [226].

With/without user commentary. The way in which different users curate content varies substantially. In most cases, users include links in their tweets to the content they have found. Sometimes, they also provide personal comments and opinions, using Twitter as both a news service and a social network [139]. For instance, @DruidSmith is a geolocation/GPS expert who, aside from linking to content from other sources, also shares his own knowledge and experience.

Human/automatic. In Twitter there is a significant amount of *news aggregators*. They collect news articles (*e.g.* from RSS feeds) and automatically post their corresponding headlines and URLs to Twitter. The majority of them post many tweets per day related to breaking news and top stories, *e.g.* @BreakingNews. A minority are focused on more specific topics, and thus constitute *topic-focused aggregators*. In all cases, they do not provide any personal contents or opinions.

For instance, @RobinGood is a widely recognised curator on the topics of media, technology and design. However, @RobinGood maintains a personal selection of blog and directories of RSS, which he updates weekly. His account distributes automatically the stories that appear in those sources. Thus, all of his tweets are generated automatically. Some news aggregators seem to be considered valuable by users, as in the case of @RobinGood who has over 17,000 followers at the time of this writing. However, whether all news aggregators provide interesting content to a topic is questionable.

In summary, there are different types of users that aggregate content. They differ in the number of topics they cover (focused/unfocused), in how much commentary they include (only URLs or URLs and comments) and in the way they post information (human/automatic). These insights allow us to make a first characterisation of news story curators. Tweeting about a story but also about many other stories that are not related (*e.g.* not the same topic or geographic region) indicates that the user is not interested in the story per se, thus, should not be considered as a curator for it. Moreover, we are not interested in finding mere news aggregators, who automatically post content from a set of sources, sometimes using automatic filters. Instead, we look for news curators, where there is human intelligence behind the choice of each individual article. As a consequence, we distinguish between human and automatic tweet creation.

Story Curators in a News Crowd

We selected two articles from our dataset and looked at their crowds, *i.e.* the users who posted these articles to their Twitter timelines. Table 11.5 lists some of these users and their characteristics. We focus on two characteristics of them: 1) whether they seem *human or automatic*, which separates news aggregators and curators, and 2) whether they seem to be *interested in the topic of the article or not*, which describes the topical focus of a user.

Table 11.5: Example of users for two news articles. We include the number of followers, tweets per day, fraction of tweets containing URLs and user mentions (“@”), the type of tweet generation and the main topic.

	#Followers	Tweets/day	Fraction		Type	Topic
			URL	@		
16 Jan 2013 – Syria allows UN to step up food aid						
@RevolutionSyria	88122	189.13	0.86	0.02	Auto.	Syria
@KenanFreeSyria	13388	9.29	0.74	0.28	Human	Syria
@UP_food	703	10.22	1.00	0.00	Auto.	Food
18 Jan 2013 – US cyclist Lance Armstrong admits to doping						
@KevinMcCallum	15287	60.15	0.18	0.77	Human	Sports
@huyanxing	3224	69.19	1.00	0.00	Auto.	Misc.
@WaseemMansour	1298	15.33	1.00	0.00	Auto.	Misc.

In Table 11.5 the first article is about the civil war in Syria. Two of the users who posted this story have several tweets related to Syria: @RevolutionSyria provides automatically generated tweets, whereas the content twitted by @KenanFreeSyria is collected by hand. We can see this by looking at the number of tweets per day for @RevolutionSyria and the fact that it has almost no user mentions (it does not engage in conversations with others). The user @UP_food is a news aggregator that apparently tweets anything containing the word “food”, but is not relevant for the developing story about Syria.

The second article is on the doping scandal of Lance Armstrong. We could detect one curator for this story, @KevinMcCallum, who routinely collects sports-related content. The other users in the crowd aggregated breaking and top news in an automatic manner (*e.g.* @huyanxing and @WaseemMansour).

11.6.2 Learning Framework

As we could see in the previous section, there are a number of issues that have to be dealt with carefully in order to develop an automatic method to detect news story curators.

We defined two tasks: the first one detects users that are interested in the given story or topics associated with the article (`UsersInterestedInStory`), and the second one evaluates whether the user is human or generates its tweets automatically (`UsersHuman`). We employ a random forest classifier after

information-gain-based feature selection, as implemented in Weka.⁷ We use standard evaluation metrics such precision, recall, and AUC, all measured after ten-fold cross validation.

We focus on identifying news story curators for world news, that is, for news crowds formed by AJE and BBC-WORLD. We do not consider BBC-UK news crowds, as such crowds focus on UK news. We know from Section 11.5.5 that UK news are of temporal nature meaning that it is less likely that users become curators of such stories.

We define two criteria to reduce the number of users under consideration (*i.e.* potential story curators). First, we examine only users with at least 1,000 followers, because users with less followers are not influential enough to play a significant role in the Twitter ecosystem [231]. Second, we consider only users who posted at least one URL related to the original news article (according to our method defined in Section 11.5).

11.6.3 Training Data

In an automatic system, we apply supervised learning to detect story curators, and thus a training set (human-provided labels) is required. We created our training data by selecting a sample of 20 news articles: 10 from AJE, and 10 from BBC-WORLD. For each news article, we sampled uniformly at random 10 users who posted the article. We then asked three volunteers to provide labels.⁸ We provided them examples and typical characteristics of the various types of news aggregator and curator (as discussed in Section 11.6.1).

For the labelling task, we showed the title of the news article and a sample of tweets of the user. We showed tweets that were posted directly after the news article, since the lifetime of some stories can be very short. We also presented the profile description and the number of followers of the user. Then, we asked our annotators to label the user as follows:

- [Q1] Please indicate whether the user is interested or an expert of the topic of the article story: Yes, maybe, no, or unknown.
- [Q2] Please indicate whether the user is a human or generates tweets automatically: Human, maybe automatic, automatic, or unknown.

⁷<http://www.cs.waikato.ac.nz/ml/weka/>

⁸All are computer science or engineering graduates with experience in using Twitter.

Table 11.6: Distributions of the human-provided labels.

Dataset	Interested?			Human or Automatic?		
	<i>n</i>	yes	not	<i>n</i>	human	automatic
AJE	63	21%	79%	71	55%	45%
BBC-WORLD	58	3%	97%	54	35%	65%

A detailed overview of the instructions can be found in the Appendix in Section A.3.2. In total, 417 labels were collected. We decided to label whether a user is interested in the topic of the news story or not, instead of asking whether the user is an expert of that topic. We assume that users that are not experts (*e.g.* eye-witnesses), but interested in the story, could reveal interesting information for journalists.

For the training set, we considered only users for which at least two annotators provided a decisive label (Yes or No, Human or Automatic). We discarded any “maybe”, “maybe automatic”, and “unknown” labels, since these users could be used neither for training nor evaluation purposes. The distribution of labels is shown in Table 11.6. While 13% of the AJE users were labelled as both interested in the topic and human, only 1.8% of them had both labels in the case of BBC-WORLD.

11.6.4 Features

Previous work including [90; 259; 271] provides us with some useful information about suitable features for the detection of curators. These include network-based features such as the number of followers of a user – shown not to be sufficient on its own as a predictor of expertise by [271] – as well as contextual features including user profile and user lists [259].

Our features try to capture three aspects of users: (1) the *visibility* of a user; (2) characteristics of the user’s tweets that might separate human from automatic users; and (3) how focused are the tweets of users with respect to the news media source.

We transformed the frequency-based values to provider-specific quantile values in the filtered dataset, since we are merging users coming from two different news providers whose audiences have different sizes, as we showed in Figure 11.1. These features are denoted by the suffix *Q* in the feature name.

Visibility. The visibility of a Twitter user is a core characteristic of a curator. There are different features that can be associated with a user visibility. This can be captured by the number of followers (`UserFollowersQ`) or the number of Twitter lists containing that user (`UserListedQ`). We remark that both features are highly correlated in our data ($r^2 = 0.96$), which is consistent with the findings of Sharma et al. [231]. However, we do not know a priori if one of the features is more important than the other in any of the two classification tasks that we attempt.

Tweeting activity. In Section 11.6.1 we described the presence of prolific automatic accounts in Twitter. Features that capture the tweeting activity of a user may reflect best the differences between human and automatic accounts. We measure the number of tweets per day (`UserTweetsDailyQ`), the fraction of tweets that contains a re-tweet mark “RT”, a URL, a user mention or a hashtag (respectively, `UserFracRetweets`, `UserFracURL`, `UserFracMention`, and `UserFracHashtag`).

Topic focus. A natural measure of topical focus is how many different articles in each dataset has this user tweeted (`UserCrowdsQ`). Additionally, as articles belong to a section in each website (*e.g.* sports, business, Europe, USA), we also define the number of distinct sections of the crowds s/he belongs to (`UserSectionsQ`).⁹

11.6.5 Results

We tried two types of models, one considering only a single input feature, and one considering all the input features.

Simple models. For the task `UserIsHuman`, a basic but effective approach is to apply a simple rule, which yields a precision and recall of 0.85 (for the human class), and an AUC of 0.81:

`UserFracURL` $\geq 0.85 \Rightarrow$ automatic, otherwise human.

This means that a news aggregator (automatic user) can be readily identified because a large fraction of its tweets contain URLs. This agrees with previous works (*e.g.* [22]) and our manual analysis of the data in Section 11.6.1.

⁹The section of an article can be extracted from the prefix of the path of the article. For instance, articles under <http://www.bbc.co.uk/news/world-latin-america-21001060> correspond to the section “Latin America” of BBC-WORLD. In websites organised in a different manner, other ways of defining sections may be necessary.

For the task `UsersInterestedInStory`, the following rule yields a precision of 0.48 (remember the classes are not balanced), a recall of 0.93 (for the `interested` class), and an AUC of 0.83:

```
UserSectionsQ >= 0.9 => not-interested, otherwise interested
```

This means that if a user does not tweet about many different sections of a news site, one can expect that the sections s/he is tweeting about relate to the topic of the story, thus, s/he is interested in the story of the given article. However, it is not always the case as we can see in Table 11.5: the curator `@UP_food` collects tweets around the topic “food”, which is not relevant for the story about Syria.

Complex models. More complex models based on the random forest classifier perform better for the `UsersHuman` task (AUC 0.93 vs. AUC of 0.85 for the single-feature model). As expected, all features related to the tweeting activity (`UserFracRetweets`, `UserFracURL`, `UserFracMention`, and `UserFracHashtag`) are the most important features for this model. Adding more features to the model for the `UsersInterestedInStory` task also yields an improvement in performance when comparing with the single-feature model (AUC 0.90 vs. AUC 0.83), and might also improve the robustness of the prediction. Given a large class imbalance for the `UsersInterestedInStory` task, we applied asymmetric misclassification costs. Specifically, false negatives (classifying an interested user as not interested) were considered 5 times more costly than false positives; values close to this number did not change substantially the obtained results.

All results are summarised in Table 11.7. Overall, we were able to demonstrate that the considered features can be used to automatically find news (story) curators in transient news crowds. Note that, as shown in Section 11.3 (Figure 11.1), there is a difference in the sizes of the audiences of the news providers, `BBC-WORLD` and `AJE`; nonetheless, we could identify story news curators for both.

11.6.6 Precision-oriented Evaluation

We also compared our method with two baseline approaches: (1) select the users with the largest number of followers among the candidates, and (2) select the users with the largest number of content detected as related to the original one (using the method of Section 11.5).

Table 11.7: Evaluation of models for the `UserIsHuman` and `UserIsInterestedInStory` tasks.

	Precision	Recall	AUC
<code>automatic</code>	0.88	0.84	0.93
<code>human</code>	0.82	0.86	0.93
<code>interested</code>	0.95	0.92	0.90
<code>not-interested</code>	0.53	0.67	0.90

Data. We selected a sample of 20 news articles that had at least one curator detected using the model that uses all the features with a confidence value ≥ 0.75 . For comparison, we extracted for each article the same number of possible curators using the other two approaches. Then, we merged the results together without providing which system identified which curator. We asked the same three assessors to evaluate the results using the question (Q1) of Section 11.6.3.

Results. We collected about 210 labels for 70 units. The assessors labelled 71% of the users as *not interested*, 6% as *interested*, and 15% as *maybe interested*. We merged the labels *yes* and *maybe*, and we considered only users for which at least two assessors had the same label. As a consequence an unequal number of labels per approach is given. The worst performance was obtained by the follower-based approach ($2/18 = 11\%$): only two users with a high number of followers were labelled as curators and 18 users with a high number of followers were not curators. A better performance was obtained by the automatic detection of related content ($5/20 = 25\%$), but our approach outperformed the other two ($6/16 = 37.5\%$).

11.7 Discussion

This chapter showed how inter-site engagement, more precisely, the fact that users also visit Twitter for their daily news consumption, can be used to help journalists and news editors of mainstream media outlets in rapidly acquiring follow-up articles and information about the stories they are writing about. We propose to do so by leveraging *transient news crowds*, which are loosely-coupled groups that appear in Twitter around a particular news item.

That a user posts a news article to a microblogging site may seem a trivial action, and, indeed, in terms of individuals, not much can be read from this event. However, when we consider users in aggregate, we uncover a noisy

yet usable signal that can lead to new insights and to the discovery of new related content. This content is of particular interest for news providers, because we have shown that providing related content to a news story can have a positive impact on the user engagement with the news provider (see Chapter 10).

We have observed that after users tweet a news article, their subsequent tweets are correlated during a brief period of time. We have shown that such correlation is weak but significant, in terms of, for instance, reflecting the similarity between the articles that originate a crowd. We have also shown that just as the majority of crowds simply disperse over time, parts of some crowds come together again around new newsworthy events related to the original story. This is in accordance with the results of Chapter 10 where we could see that if users are interested in a story, they decide to focus on it by reading several articles about the story.

As an application of this observation, we have designed and validated experimentally a method for uncovering related content to a news story. This method can be used to build a practical system in which a journalist can be presented with a selection of news and other information that are often related to the one s/he originally authored.

In the second part of this work, we have defined and modelled a class of users, news story curators, that has the potential to play an important role in the news ecosystem, particularly for journalists, editors, and readers. We have found that finding news story curators is a challenging task. First, there is a large amount of automatic activity on Twitter, and some of these news aggregators are actually considered by some users to be good curators. Second, posting a link in Twitter may or may not reflect a long-standing interest on the subject of the link.

In our approach, we have automatically found news story curators. A key aspect of that system is being able to assess how spread are the interests of a user. This matched our intuitions in the sense that the more diverse a user's interests are, the less likely that person is to be a good news story curator. Next, we have tackled this problem by trying to separate automatically-operated accounts from manually-operated ones, showing that while simple rules can be somewhat effective, combining different aspects of the information available about a user can yield better results.

A fundamental concern for journalists is how to find reliable information sources on a given story. These sources are usually either (i) primary sources

that can provide authoritative information by themselves, such as eye witness of a developing crisis, or (ii) secondary sources that are proficient at aggregating information from primary sources, such as specialised news curators. The study of transient news crowds can help in both tasks.

Limitations. Our results have some limitations. We did not attempt to analyse how valuable are the related content and news story curators for journalists and news editors. Related content can be of different type such as news articles, and pictures or videos from eye-witnesses, and also the quality of the content can differ. In addition, we did not differentiate between types of news story curators, such as opinion leaders, who are news curators that tweet directly from the news website (using the “share” buttons and not retweeting). Opinion leaders are considered to mediate between mass media and the public in the so called two-step flow of information [274].

Considering the fleeting nature of news, our approach might not be applicable for linking to related content on the news article that created the crowd; it detects related content at the end earliest after 6h. However, the information can be used by journalists to write new articles covering novel *beats* related to the story, and such articles (and others) can link to the related content; thereby also taking into account the shifts of interest of the audience.

Finally, we did not analyse how the definition of “relatedness” affects our method. In Section 11.5.1, we could see that it is possible to detect more related content using a less restrictive definition, as in this case the “Algerian hostage” would relate to the “Mali conflict”. However, we did not attempt to study how the strength of “relatedness” affects the precision and accuracy of our method.



Conclusions and Future Work

On the Internet, user engagement has become a key factor for the evaluation of a website success. One standard approach to measure engagement, and the focus of this dissertation, is through the usage of online behaviour metrics. This chapter discusses the results and conclusions of the thesis and proposes research directions for future work. We start by summarising the key findings of each chapter, which are presented in Figure 12.1.

12.1 Summary

The fundamentals part (Part II) of the thesis was concerned with the study of different aspects of user online behaviour. Based on the insights gained from this analysis, new metrics to study user engagement were defined. All these metrics aim to support providers by understanding their engaged audience better.

Chapter 4 studied how existing engagement metrics reflect the engagement of users with a site. We referred to this type of engagement as *site engagement*. Three types of metrics, capturing the popularity, activity, and loyalty of users with a site, are widely used to analyse user engagement. It has been shown that sites differ in their engagement, and that this does not necessarily mean that one site is more engaging than another; engagement depends on the site at hand. To characterise these differences, we defined patterns of engagement that provided us with different but complementary insights about how users engage with sites.

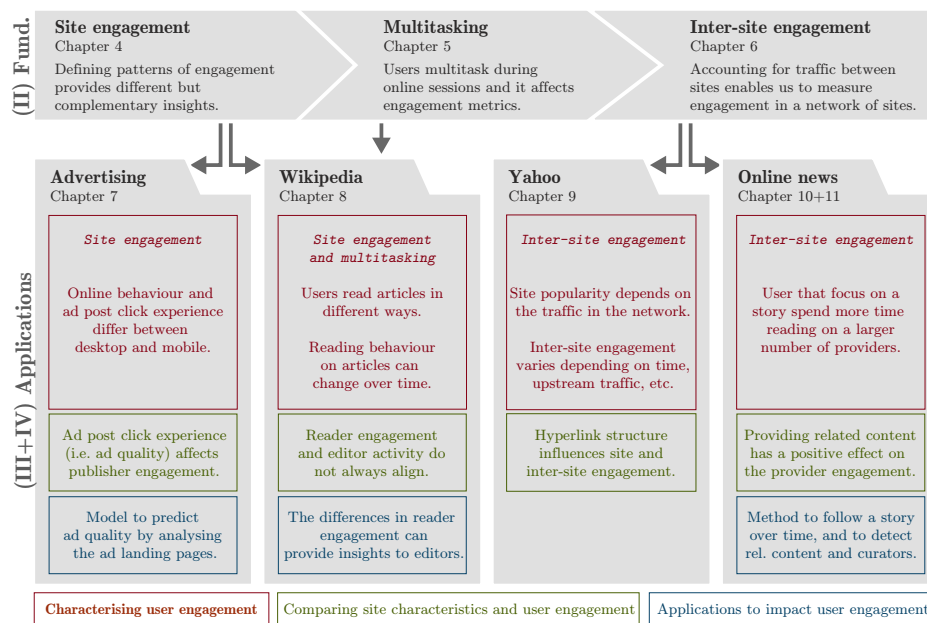


Figure 12.1: Key findings in each chapter of the dissertation.

Chapter 5 explored a so far under-considered aspect of user online behaviour – *online multitasking*. We showed that users frequently visit and revisit sites during their online sessions and that such behaviour affects engagement metrics that measure the activity of users on a site. At the end of the chapter, we defined and evaluated new engagement metrics that reflect users’ multitasking behaviour with sites.

How engagement can be measured within a network of sites, which we referred to as *inter-site engagement*, was studied in Chapter 6. The fact that users multitask within online sessions, and that many providers offer not only one but many different services, makes it increasingly important to measure engagement across sites. We therefore modelled sites and the traffic between them as networks, and defined and evaluated new metrics that account for the traffic in such networks.

The insights gained in the fundamentals part of the thesis were then used as the basis to carry out four case studies. Each case study was in addition concerned with how site characteristics influence user engagement, and, based on the findings, the development of approaches that have the potential to increase user engagement.

Application I consists of two case studies, which were about site engagement and multitasking. Chapter 7 focused on the relationship between advertising quality and the engagement with the site that serves the ads. The quality of the ads was assessed through users' experience with the ad landing page, namely their ad post-click experience. We could show that ad quality affects the engagement with the publisher site, but that the perceived quality is device-dependent (desktop vs. mobile). Based on these observations we developed a model that predicts ad quality on mobile devices.

Chapter 8 studied the relationship between readers and editors in Wikipedia. Using site engagement and multitasking metrics, we showed that users read articles in Wikipedia in different ways, and that the reading behaviour on an article can change over time. It was also observed that the engagement of readers and the activity of editors do not always align. Finally, we discussed how information about users' reading behaviour can assist the Wikipedia editor community in their editing tasks.

The last two case studies constitute the second applications part of the thesis (Application II). The studies were concerned with inter-site engagement. Chapter 9 characterised inter-site engagement in a network of sites offered by the same provider, in our case Yahoo. We found that inter-site engagement depends on various factors such as user loyalty, and the type of upstream traffic. We also saw that there is a strong network effect, that is, the popularity of sites depends on the traffic between them, and vice versa. The results of the chapter also suggested that the hyperlink structure of the network influences the traffic, and hence site and inter-site engagement.

Finally, Chapters 10 and 11 explored how users find and consume content related to a news story. The first chapter studied how users focus on stories within and across news sites. Story-focused reading leads to a higher site and inter-site engagement – users spend more time reading and engage with a larger number of news sites. Moreover, news sites can keep users engaged by linking to related content on their article pages. Motivated by this, we then studied how users follow stories in Twitter, and we developed an approach that rapidly detects content and curators related to that story.

To summarise, the contributions of the thesis are as follows. First, user online behaviour was studied and new metrics to measure engagement were defined. Second, four case studies were performed where existing and new metrics were applied, and a deeper understanding was attained about engagement and which factors influence it. In the following section we discuss in detail the insights gained in the thesis.

12.2 Research Results

We answer the research questions posed at the beginning of the thesis (Section 1.1), and discuss the corresponding results in detail.

Q1: Patterns of User Engagement

How does user engagement differ between sites and how can we measure and characterise such differences? What should be taken into consideration when measuring user engagement?

The first questions were answered in the fundamentals part of the thesis (Part II). We showed that user engagement differs with respect to how users engage with sites (Chapter 4), how they multitask with sites during online sessions (Chapter 5), and how they engage with a network of sites (Chapter 6). These differences do not mean that the engagement of some sites is higher compared to others, it is simply different.

To characterise these differences, we determined patterns of user engagement by clustering the sites based on their site engagement, multitasking, and inter-site engagement characteristics. Each pattern reflects the engagement features of a certain group of sites. The same method was used in Chapter 8 to characterise the reading behaviour of users in Wikipedia. Since a detailed description of the patterns of sites is given in Chapters 4 to 6, we only discuss two examples here.

Using site engagement metrics, we have shown that news sites are engaging when users are visiting them frequently and spending a lot of time on them during each visit (high loyalty and activity). However, the multitasking patterns differ between news sites. Some news sites exhibit a single-task-oriented browsing activity – the focus of users is on the news site within an online session. Other news sites follow a multitask-oriented browsing pattern, that is, users visit the site many times within an online session with short breaks in-between the visits. The multitasking effect for these sites is significant – the dwell time per visit is low, but the aggregated dwell time (over the online session) is high. Finally, we studied how users engage with news sites within a provider network. We could see that news sites are well connected to other sites in the network. Although users dwell long on news sites, it is very likely that they also continue browsing to other (probably news) sites in the network.

In contrast, for shopping sites it is *not* important that users visit the sites frequently (*i.e.* shopping is not a daily activity), what matters is that users dwell long during each visit (low loyalty, high activity). Moreover, many shopping sites follow the continuous multitasking pattern, that is, although users spend a lot of time on the site during each visit, they are continuously returning to the site within an online session, even after a longer time of absence. Users might read product reviews and visit other shopping sites in the meantime – they want to select the best product with the best price.

To summarise, patterns of site engagement, multitasking, and inter-site engagement demonstrate that differences in engagement are a consequence of having sites that provide different services with different characteristics – engagement is service-dependent.

This does, however, not imply that all sites are to the same extent engaging. There is also variance between the features (*e.g.* dwell time, return rate) of sites that follow the same pattern (cluster) reflecting differences in the engagement of sites that exhibit the same basic engagement characteristics. In addition, some sites might follow the “wrong” pattern. For instance, it is questionable whether a shopping site is engaging when following the “quick task” pattern (Section 5.9), which is characterised by a low dwell time, and users visit the site only once during an online session. However, the aim of the thesis was not to identify which sites are more engaging than others, but to provide a deeper understanding of how users engage with sites, through, for example, the development of new metrics.

When measuring user engagement, not only the characteristics of a service, but also other aspects should be taken into consideration. Chapters 4 and 9 showed that site and inter-site engagement also depend on the loyalty of users, temporal aspects, and the type of upstream traffic. Indeed, the engagement is higher during the weekend for leisure-oriented websites (*e.g.* shopping), whereas it is higher on weekdays for goal-oriented websites (*e.g.* mail). Moreover, we could see in Chapter 8 that the engagement with Wikipedia articles can change over time, and that mainly exogenous factors are responsible for these changes (*e.g.* a person becomes famous).

Chapter 7 demonstrated that engagement is device-dependent. Users can be highly engaged to an ad when using a desktop device (high post-click experience), but this experience can be different on a mobile device (low post-click experience). Finally, we could see that users have a limited amount of time to spend online. In a provider network of sites, this may imply that if users spend more time on one site, they will spend less time on another

site (Chapter 9). If users have to visit many sites, they are likely to do so quickly, maybe only skimming the content of the pages, to not exceed their available time (Chapters 5 and 10).

To summarise, various aspects should be taken into account when measuring user engagement, since engagement differs between sites for many reasons. Defining patterns of user engagement enabled us to characterise such differences in a clear manner.

Q2: Online Multitasking

Does online multitasking affect engagement metrics that capture users' browsing activity on a site? Is it possible to define metrics that characterise multitasking during online sessions and do these metrics provide new insights compared to standard engagement metrics?

Chapter 5 studied online multitasking and how it affects the measurements of user engagement. Nowadays, users can visit a large range of sites to perform various tasks, and, in addition, GUI elements such as browser tabs allow them to easily perform tasks in parallel. We have shown that online multitasking exists, as many sites are visited and revisited within an online session. We also demonstrated that multitasking influences the way users access sites and that this affects engagement metrics that capture the browsing activity on sites. As a result, studying engagement *at* visit-level can be misleading. Instead, the browsing activity should be measured *at* session level, or on a daily basis (*e.g.* dwell time per day).

Multitasking also implies that leaving a site does not necessarily entail less engagement, as users often return to the site later on within the online session (Chapter 5). This was further demonstrated in Chapters 6 and 9, where we could show that users frequently switch between the sites in a provider network and that they often leave but also return to the provider network. This should be considered by online providers who often aim to keep users as long as possible on their sites. Indeed, in Chapter 10, we could see that encouraging users to leave a news site, by providing links to story-related content on other sites, does not harm the engagement with the news site. Instead, we observed that users spend the same time reading on the news site, and that they come back sooner to start a new reading session.

Motivated by these observations, we defined five metrics that characterise the multitasking behaviour with a site. We showed the value of these metrics in two ways. First, we analysed the online activity on 760 different sites us-

ing both multitasking and engagement metrics (Chapter 5), and we demonstrated that multitasking metrics provide new insights about how users engage with a site. We then analysed how sites differ in their multitasking behaviour by generating five multitasking patterns. The first two patterns characterised sites with a single-task-oriented behaviour (no multitasking), and the other three patterns described sites that exhibit a multitask-oriented behaviour.

Second, we used some of these metrics in Chapter 8 to analyse the reading behaviour of users on Wikipedia. The multitasking metrics provided insights about users' reading behaviour on Wikipedia articles that could not have been identified with standard engagement metrics. For instance, there were sessions in which users explore a topic in Wikipedia by reading many articles related to it. During such sessions users only passed through the focal article, or they used it as a basis to navigate to other articles on the same topics.

Overall, we have demonstrated that users multitask within their online sessions, and that multitasking can affect how we interpret standard engagement metrics. We defined new metrics that characterise the multitasking behaviour with a site (or page), and showed that such metrics provide new insights about user engagement.

Q3: Inter-site Engagement

*How can we measure user engagement with respect to a network of sites?
How does this enhance our understanding of engagement?*

We started answering these questions in Chapter 6 by modelling sites and the user traffic between them as a network. We defined various inter-site metrics at network- and node-level that characterise the engagement, respectively, with a network and with the sites in a network. In contrast to site engagement and multitasking metrics, these metrics account for the traffic between sites. We then showed how these metrics provide new insights about user engagement.

Chapters 6 and 9 used the metrics to analyse the engagement of users in provider networks, that is, networks that consist of sites (services) belonging to the same provider. We defined several such networks, for example based on a subset of sites (those related to one country), or a subset of traffic (weekend traffic). We showed, for instance, that network-level metrics enable us to compare whole provider networks. We could see that some networks are engaging with respect to the time users spend on the network (high

provider engagement), but not with respect to the traffic between sites (low inter-site engagement). Similar observations could be made with the metrics at site-level. For instance, front pages have a high inter-site engagement – they are crucial for a provider network as they direct traffic to other sites. However, these sites are not engaging with respect to the time users spend on them.

Chapter 10 studied inter-site engagement with respect to online news consumption. The chapter analysed how users focus on stories they are interested in by reading several articles about it. We refer to this type of reading behaviour as story-focused news reading. Since users increasingly use more than one news provider and also other source to consume news online, studying user reading behaviour at site level leads to a limited view about story-focused reading. We therefore studied how users consume story-related articles in a network of news providers and we analysed how they find these articles. The results suggest that inter-site engagement is especially interesting in the context of story-focused news reading, as a larger number of news providers are involved when users focus on a story, and the traffic flow in the network is higher.

To summarise, accounting for the traffic between sites enables us to measure user engagement with respect to a network of sites. Studying engagement from a network perspective allows for a more comprehensive look at user engagement by also considering the relationships between sites.

Q4: Factors of Influence and Applications

How do the characteristics of a site (content and hyperlinks) affect user engagement? Can we use such dependencies to develop applications that have the potential to improve the engagement of users?

These research questions were answered in Applications I and II of the thesis (Part III and IV). We analysed how certain characteristics of sites affect engagement, and based on our findings we developed approaches that have the potential to improve engagement.

First of all, it should be noticed that the characteristics of a site are not the only factor that affects engagement. As already discussed at the beginning of the dissertation (Section 2.3), the user context is also an important factor and should to be taken into account. The results of this dissertation confirm that user engagement also depends on users' interests. We observed that readers engagement in Wikipedia strongly depends on the interests of the

users and less on the content quality (Chapter 8). We also saw that users engage (focus) with news stories they are interested in (Chapter 10).

However, we demonstrated as well that the characteristics of a site, its content quality and hyperlinks, affect user engagement. We started in Application I by analysing whether and how the quality of content influences site engagement.

Chapter 7 showed that the quality of ads affect the engagement of users to the site that serves the ads – low quality ads can lead to users accessing the site less often and clicking less on ads. Based on this observation, we developed a model that predicts ad quality by analysing their landing pages and relating these to the ad post-click experience, that is, how users engage with the ad. This model can be used to serve high quality ads in a news stream with the objective to increase revenue, but also to keep users engaged by serving them appealing advertisements.

Chapter 8 suggests that reader engagement in Wikipedia is driven by the interests of the readers. It is, however, the case that readers are also interested in articles that are of low quality. We argued that increasing the quality of such articles will lead to a higher reader engagement, and we discussed how readers' engagement in Wikipedia – their preferences and behaviours – can be used by Wikipedia's editor community to make informed decisions about improving articles. The idea is to promote positive reading experiences on articles and in doing so to encourage users to return regularly to Wikipedia.

Application II analysed how the hyperlink structure of sites influences site and inter-site engagement. Chapter 9 compared the traffic in a provider network with its hyperlink structure, and found that both are similar when comparing them with PageRank. This implies that it is likely that users will visit sites where many hyperlinks are pointing to. Moreover, hyperlinks can be used to keep users longer on a site, but also to direct them to other sites in the network, that is, to increase inter-site engagement. However, it has been shown that having many hyperlinks between sites results in users visiting more sites in the network, but also spending less time on each site. Hence, there is a relationship between site and inter-site engagement, and it may be difficult to improve both at the same time.

Finally, Chapter 10 analysed how news providers link their articles to related content published by them (or other sources) and which effect this had on the engagement. It has been shown that internal links, which point to pages of the same news provider, encourage users to stay longer on the news site, and that they also have a positive effect on users' long-term engagement –

users return earlier to the site to start a new reading session. Interestingly, even links to content from other sites, *i.e.* links that promote inter-site engagement, do not harm the engagement with the site. Users acknowledge such links as they probably provide valuable information about the story, and therefore they do not engage less in terms of dwell time and even return earlier (the next reading session). However, this does not imply that news sites should just provide such links, but the right ones. We demonstrated as well that the right type, position, and amount of links are essential to provide a positive reading experience.

However, a so far not well-explored problem is how news providers, their journalists and editors, can find related content in which their audience is interested. We tackled this problem as follows. Nowadays, the audience of a news provider does not only engage with the news site by visiting the site frequently and spending time on it, but also through social media sites by sharing the news articles published by the site. We used this aspect of inter-site engagement and analysed the audiences of news articles in Twitter (Chapter 11). We followed these audiences over time and showed that parts of the audience share again content related to the story. Based on this observation, we developed an approach that rapidly detects related content to and the curators of a story. News sites can link their articles to such content, which in return might have an impact on the user engagement with the site.

Overall, site characteristics such as content quality and hyperlinks influence site and inter-site engagement. Being aware of such dependencies can help to develop applications that might have an impact on engagement.

In summary, the thesis aimed to provide a deeper understanding of user engagement. We started by studying how engagement differs between sites and which aspects should be taken into account when measuring it (Q1). We then incorporated new aspects of user engagement. We realised that users multitask and that such online behaviour should be taken into account when measuring user engagement (Q2). Multitasking implies that users access and re-access many sites within an online session - either to perform the same or totally unrelated tasks. This implies that instead of measuring engagement with a site, we can also measure the engagement with a network of sites by accounting for the traffic between the sites (Q3). Finally, since it is the aim of providers to design engaging experiences, we also investigated which factors influence engagement, and whether we can use our observations to develop applications that can have the potential to increase engagement.

12.3 Future Work

Now, we discuss possible directions for future research.

Case Studies on other Applications

We plan to further evaluate the applicability of our multitasking and inter-site metrics on other applications. For instance, an important activity on the Web that we did not consider is online shopping. Users invest time by comparing offers and reading reviews about a product, thereby accessing and re-accessing many sites within an online session. Our metrics can be used to study this behaviour. We should also take into account that shopping tasks can be performed over a longer time (*i.e.* across sessions), and that users visit a site for different purposes (reading reviews vs. purchasing an item).

Inter-site metrics can be also used to study engagement within a site by modelling pages (nodes) and the traffic (edges) between them as a network. For instance, the downstream engagement metric can inform news or shopping providers about which news articles and shopping items are important in driving users to other articles or items.

Factors of Influence

The work about the relationship between site characteristics and engagement can be extended. Other aspects of the site content and its effects on engagement could be studied, such as its novelty, completeness, and serendipity. It is also essential to gain a deeper understanding about the relationship between hyperlinks and inter-site engagement, especially across sites from different providers. We could show that providing links that direct users to other sites does not necessarily harm the engagement with the site. Finding sites that are anyway visited from the same audience and interlinking them, might have a positive impact on the engagement with all these sites. To a certain extent, a first step in this direction is the integration of social media platforms (the “share” buttons) into news sites.

Self-reported and Physiological Methods

In Section 2.2, we mentioned that user engagement is also measured through self-reported and physiological methods. Since the focus of the thesis was on measuring user engagement through online behaviour methods, such methods have not been considered. However, these methods can enhance our

understanding of users' multitasking habits and how they engage with a network of sites. In addition, these methods can provide valuable information about the meaning of our metrics and their interpretation.

For instance, eye and mouse tracking can be used in the context of multitasking and inter-site engagement to better understand how users switch between sites. In addition, online questionnaires can be used as a feedback mechanism to understand whether a certain online behaviour can be considered as engaging or not (*e.g.* users accessing many articles on Wikipedia can also be a sign of frustration, since they cannot find the information they are looking for).

User Engagement across Devices

Only a small part of the dissertation was concerned with user online behaviour on mobile devices. However, users increasingly use their mobile phones and tablets to perform various tasks on the Web. Two aspects of users' mobile engagement are especially interesting.

First, the analysis of inter-site engagement in a provider network across devices. We assume that studying user online behaviour on only one device leads to an incomplete picture about users' engagement. For example, some users may only read news using a desktop device. However, when taken into account also other devices, we might observe that the same users read emails using a mobile phone, watch videos using a tablet, *etc.* (high engagement across devices).

Second, a study about how users engage with several mobile applications to perform a task, thereby taking into consideration the location-dependent context (*e.g.* using a map service and websites containing reviews when looking for a restaurant). In general, we assume that users switch less between sites on a mobile device, given that the interface is smaller, and also tab switching is more complicated than on desktop devices. However, this also leads to a higher accuracy and amount of data as users interact with the Web in a more controlled environment.

Bibliography

Each reference indicates the pages where it appears.

- [1] Mani Abrol, Uma Mahadevan, Kenneth McCracken, Rajat Mukherjee, and Prabhakar Raghavan. Social networks for enterprise webs. In *Proc. Conference on World Wide Web, WWW*, 2002. 207
- [2] B Thomas Adler, Krishnendu Chatterjee, Luca De Alfaro, Marco Faella, Ian Pye, and Vishwanath Raman. Assigning trust to wikipedia content. In *Proc. Symposium on Wikis and Open Collaboration, Wikisym*, page 26. ACM, 2008. 19, 126
- [3] Atul Adya, Paramvir Bahl, and Lili Qiu. Analyzing the browse patterns of mobile clients. In *Proc. Workshop on Internet Measurement, SIGCOMM*, pages 189–194. ACM, 2001. 101, 102
- [4] Deepak Agarwal, Bee-Chung Chen, and Xuanhui Wang. Multi-faceted ranking of news articles using post-read actions. In *Proc. Conference on Information and Knowledge Management, CIKM*, pages 694–703. ACM, 2012. 206
- [5] Everaldo Aguiar, Saurabh Nagrecha, and Nitesh V Chawla. Predicting online video engagement using clickstreams. arXiv preprint arXiv:1405.5147, 2014. 16
- [6] Kamal Ali and Mark Scarr. Robust methodologies for modeling web click distributions. In *Proc. Conference on World Wide Web, WWW*, pages 511–520. ACM, 2007. 183
- [7] Aris Anagnostopoulos, Ravi Kumar, and Mohammad Mahdian. Influence and correlation in social networks. In *Proc. Conference on*

- Knowledge Discovery and Data Mining*, SIGKDD, pages 7–15. ACM, 2008. 180
- [8] Judd Antin and Coye Cheshire. Readers are not free-riders: reading as a form of participation on wikipedia. In *Proc. Conference on Computer Supported Cooperative Work*, CSCW, pages 127–130. ACM, 2010. 126
- [9] Ioannis Arapakis, Mounia Lalmas, B Barla Cambazoglu, Mari-Carmen Marcos, and Joemon M Jose. User engagement in online news: Under the scope of sentiment, interest, affect, and gaze. *Journal of the Association for Information Science and Technology*, 2014. 15
- [10] Ioannis Arapakis, Mounia Lalmas, Hakan Ceylan, and Pinar Donmez. Automatically embedding newsworthy links to articles: From implementation to evaluation. *Journal of the Association for Information Science and Technology*, 65(1):129–145, 2014. 192, 194
- [11] Ioannis Arapakis, Mounia Lalmas, and George Valkanas. Understanding within-content engagement through pattern analysis of mouse gestures. In *Proc. Conference on Information and Knowledge Management*, CIKM. ACM, 2014. 15
- [12] Jaime Arguello, Fernando Diaz, and Jamie Callan. Learning to aggregate vertical results into web search results. In *Proc. Conference on Information and Knowledge Management*, CIKM, pages 201–210. ACM, 2011. 174
- [13] Simon Attfield, Gabriella Kazai, Mounia Lalmas, and Benjamin Piwowarski. Towards a science of user engagement (position paper). In *Proc. Workshop on User Modelling for Web Applications*, WSDM, 2011. 1, 14, 17, 18
- [14] Javad Azimi, Ruofei Zhang, Yang Zhou, Vidhya Navalpakkam, Jianchang Mao, and Xiaoli Fern. Visual appearance of display ads and its effect on click through rate. In *Proc. Conference on Information and Knowledge Management*, CIKM, pages 495–504. ACM, 2012. 102, 103
- [15] Ricardo A. Baeza-Yates and Berthier Ribeiro-Neto. *Modern information retrieval*, volume 463. ACM Press New York, 1999. 178
- [16] Firdaus Banhawi, Nazlena Mohamad Ali, and Hairuliza Mohd Judi. User engagement attributes and levels in facebook. *Journal of Theoretical and Applied Information Technology*, 41(1):11–19, 2012. 15, 153
- [17] Albert-László Barabási, Réka Albert, and Hawoong Jeong. Scale-free characteristics of random networks: the topology of the world-wide web. *Physica A: Statistical Mechanics and its Applications*, 281(1):

- 69–77, 2000. 77
- [18] Thomas Beauvisage. The dynamics of personal territories on the web. In *Proc. Conference on Hypertext and Hypermedia*, HT, pages 25–34. ACM, 2009. 33, 76
- [19] Hila Becker, Andrei Broder, Evgeniy Gabrilovich, Vanja Josifovski, and Bo Pang. What happens after an ad click?: quantifying the impact of landing pages in web advertising. In *Proc. Conference on Knowledge Discovery and Data Mining*, SIGKDD, pages 57–66. ACM, 2009. 103, 104, 114
- [20] Hila Becker, Andrei Broder, Evgeniy Gabrilovich, Vanja Josifovski, and Bo Pang. Context transfer in search advertising. In *Proc. Conference on Research and Development in Information Retrieval*, SIGIR, pages 656–657. ACM, 2009. 104, 114
- [21] Fabrício Benevenuto, Tiago Rodrigues, Meeyoung Cha, and Virgílio Almeida. Characterizing user behavior in online social networks. In *Proc. Conference on Internet measurement*, SIGCOMM, pages 49–62. ACM, 2009. 2, 17
- [22] Fabricio Benevenuto, Gabriel Magno, Tiago Rodrigues, and Virgilio Almeida. Detecting spammers on twitter. In *Proc. Conference on Collaboration, Electronic Messaging, Anti-Abuse and Spam*, volume 6 of *CEAS*, page 12, 2010. 209, 231
- [23] Omar Besbes, Yonatan Gur, and Assaf Zeevi. Optimization in online content recommendation services: Beyond click-through-rates. *Columbia Business School Research Paper*, (14-33), 2014. 177
- [24] Rohit Bhargava. Manifesto for the content curator: The next big social media job of the future. *Influential Marketing Blog (IMG)*, 2009. 226
- [25] Geoffray Bonnin, Armelle Brun, and Anne Boyer. Taking into account tabbed browsing in predictive web usage mining. In *Proc. Conference on Social Eco-Informatics*, SOTICS, pages 49–54, 2011. 47
- [26] Kate E Boudreau. *Mobile Advertising and its Acceptance by American Consumers*. PhD thesis, Roger Williams University, 2013. 102
- [27] Christos Bouras and Vassilis Tsogkas. Assigning web news to clusters. In *Proc. Conference on Internet and Web Applications and Services*, ICIW, pages 1–6. IEEE, 2010. 176
- [28] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001. 116
- [29] Igor Brigadir, Derek Greene, and Pádraig Cunningham. A system for twitter user list curation. In *Proc. Conference on Recommender*

- systems*, RecSys, pages 293–294. ACM, 2012. 207
- [30] Andrei Broder, Marcus Fontoura, Vanja Josifovski, and Lance Riedel. A semantic approach to contextual advertising. In *Proc. Conference on Research and Development in Information Retrieval*, SIGIR, pages 559–566. ACM, 2007. 104
- [31] Andrei Broder, Massimiliano Ciaramita, Marcus Fontoura, Evgeniy Gabrilovich, Vanja Josifovski, Donald Metzler, Vanessa Murdock, and Vassilis Plachouras. To swing or not to swing: learning when (not) to advertise. In *Proc. Conference on Information and Knowledge Management*, CIKM, pages 1003–1012. ACM, 2008. 103
- [32] Randolph E Bucklin and Catarina Sismeiro. A model of web site browsing behavior estimated on clickstream data. *Journal of Marketing Research*, 40(3):249–267, 2003. 2, 16, 17, 76
- [33] Georg Buscher, Susan T Dumais, and Edward Cutrell. The good, the bad, and the random: an eye-tracking study of ad quality in web search. In *Proc. Conference on Research and Development in Information Retrieval*, SIGIR, pages 42–49. ACM, 2010. 19
- [34] Andy Carvin. Distant witness. CUNY Journalism Press, 2013. 204, 226
- [35] Lara D Catledge and James E Pitkow. Characterizing browsing strategies in the world-wide web. *Computer Networks and ISDN systems*, 27(6):1065–1073, 1995. 22, 176
- [36] The PEW Research Center. Understanding the participatory news consumer. http://www.pewinternet.org/~media/Files/Reports/2010/PIP_Understanding_the_Participatory_News_Consumer.pdf, 2010. 3, 77, 174
- [37] The PEW Research Center. In changing news landscape, even television is vulnerable. <http://www.people-press.org/files/legacy-pdf/2012NewsConsumptionReport.pdf>, 2012. 77, 173, 174, 204
- [38] Hakan Ceylan, Ioannis Arapakis, Pinar Donmez, and Mounia Lalmas. Automatically embedding newsworthy links to articles. In *Proc. Conference on Information and Knowledge Management*, CIKM, pages 1502–1506. ACM, 2012. 19, 176, 177
- [39] Peter McFaul Chapman. *Models of engagement: Intrinsically motivated interaction with multimedia learning software*. PhD thesis, University of Waterloo, 1997. 13
- [40] Jilin Chen, Rowan Nairn, Les Nelson, Michael Bernstein, and Ed Chi. Short and tweet: experiments on recommending content from infor-

- mation streams. In *Proc. Conference on Human Factors in Computing Systems*, SIGCHI, pages 1185–1194. ACM, 2010. 204
- [41] Wei Chen, Yajun Wang, and Siyu Yang. Efficient influence maximization in social networks. In *Proc. Conference on Knowledge Discovery and Data Mining*, SIGKDD, pages 199–208. ACM, 2009. 77
- [42] Suqi Cheng, Huawei Shen, Junming Huang, Guoqing Zhang, and Xueqi Cheng. Staticgreedy: solving the scalability-accuracy dilemma in influence maximization. In *Proc. Conference on Information and Knowledge Management*, CIKM, pages 509–518. ACM, 2013. 77
- [43] Flavio Chierichetti, Ravi Kumar, and Andrew Tomkins. Stochastic models for tabbed browsing. In *Proc. Conference on World Wide Web*, WWW, pages 241–250. ACM, 2010. 47
- [44] Anna Chmiel, Kamila Kowalska, and Janusz A Hołyst. Scaling of human behavior during portal browsing. *Physical Review E*, 80(6):066122, 2009. 76, 77
- [45] Yejin Choi, Marcus Fontoura, Evgeniy Gabrilovich, Vanja Josifovski, Mauricio Mediano, and Bo Pang. Using landing pages for sponsored search ad selection. In *Proc. Conference on World Wide Web*, WWW, pages 251–260. ACM, 2010. 102, 104, 114
- [46] Mark Claypool, Phong Le, Makoto Wased, and David Brown. Implicit interest indicators. In *Proc. Conference on Intelligent User Interfaces*, IUI, pages 33–40. ACM, 2001. 16
- [47] Robert B Cleveland, William S Cleveland, Jean E McRae, and Irma Terpenning. Stl: A seasonal-trend decomposition procedure based on loess. *Journal of Official Statistics*, 6(1):3–73, 1990. 34
- [48] Brian Clifton. *Advanced Web Metrics with Google Analytics*. SYBEX Inc., 2nd edition, 2010. 2, 4
- [49] Andy Cockburn and Bruce McKenzie. What do web users do? an empirical analysis of web use. *Journal of Human-Computer Studies*, 54(6):903–922, 2001. 76
- [50] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995. 114
- [51] Dan Cosley, Dan Frankowski, Loren Terveen, and John Riedl. Suggestbot: using intelligent task routing to help people find work in wikipedia. In *Proc. Conference on Intelligent User Interfaces*, IUI, pages 32–41. ACM, 2007. 146
- [52] Nick Craswell, David Hawking, Anne-Marie Vercoustre, and Peter Wilkins. P@ noptic expert: Searching for experts not just for docu-

- ments. In *Proc. Australasian Conference on World Wide Web (Poster)*, AusWeb, 2001. 207
- [53] Kevin Crowston, Hala Annabi, James Howison, and Chengetai Masango. Towards a portfolio of floss project success measures. In *Proc. Workshop on Open Source Software Engineering*, ICSE. ACM, 2004. 126
- [54] Mihaly Csikszentmihalyi and Mihaly Csikzentmihaly. *Flow: The psychology of optimal experience*, volume 41. HarperPerennial New York, 1991. 13
- [55] Dianne Cyr. Modeling web site design across cultures: relationships to trust, satisfaction, and e-loyalty. *Journal of Management Information Systems*, 24(4):47–72, 2008. 18
- [56] Peter J Danaher, Guy W Mullarkey, and Skander Essegaier. Factors affecting web site visit duration: a cross-domain analysis. *Journal of Marketing Research*, 43(2):182–194, 2006. 18, 19
- [57] Abhinandan S Das, Mayur Datar, Ashutosh Garg, and Shyam Rajaram. Google news personalization: scalable online collaborative filtering. In *Proc. Conference on World Wide Web*, WWW, pages 271–280. ACM, 2007. 175
- [58] Antonella De Angeli, Alistair Sutcliffe, and Jan Hartmann. Interaction, usability and aesthetics: what influences users’ preferences? In *Proc. Conference on Designing Interactive systems*, pages 271–280. ACM, 2006. 18
- [59] Gianmarco De Francisci Morales, Aristides Gionis, and Claudio Lucchese. From chatter to headlines: harnessing the real-time web for personalized news recommendation. In *Proc. Conference on Web Search and Data Mining*, WSDM, pages 153–162. ACM, 2012. 206
- [60] Paul B de Laat. Navigating between chaos and bureaucracy: Backgrounding trust in open-content communities. In *Social Informatics*, pages 543–557. Springer, 2012. 19
- [61] Juliette De Maeyer. Hyperlinks and journalism: where do they connect? In *Proc. Future of Journalism Conference*, 2011. 177
- [62] Marco de Sa, Vidhya Navalpakkam, and Elizabeth F Churchill. Mobile advertising: evaluating the effects of animation, user and content relevance. In *Proc. Conference on Human Factors in Computing Systems*, SIGCHI, pages 2487–2496. ACM, 2013. 102
- [63] Chrysanthos Dellarocas, Zsolt Katona, and William Rand. Media, aggregators, and the link economy: Strategic hyperlink formation in

- content networks. *Management Science*, 59(10):2360–2379, 2013. 176, 177
- [64] Hongbo Deng, Irwin King, and Michael R Lyu. Formal models for expert finding on dblp bibliography data. In *Proc. Conference on Data Mining, ICDM*, pages 163–172. IEEE, 2008. 207
- [65] Inderjit Dhillon, Yuqiang Guan, and Brian Kulis. A unified view of kernel k-means, spectral clustering and graph cuts. Technical report, Computer Science Department, University of Texas at Austin, 2004. 36
- [66] Nicholas Diakopoulos, Munmun De Choudhury, and Mor Naaman. Finding and assessing social media information sources in the context of journalism. In *Proc. Conference on Human Factors in Computing Systems, SIGCHI*, pages 2451–2460. ACM, 2012. 203
- [67] Florin Dobrian, Vyas Sekar, Asad Awan, Ion Stoica, Dilip Joseph, Aditya Ganjam, Jibin Zhan, and Hui Zhang. Understanding the impact of video quality on user engagement. *ACM SIGCOMM Computer Communication Review*, 41(4):362–373, 2011. 4, 16, 153
- [68] Yellowlees Douglas and Andrew Hargadon. The pleasure principle: immersion, engagement, flow. In *Proceedings of the eleventh ACM on Hypertext and hypermedia*, pages 153–160. ACM, 2000. 13
- [69] Daniel Drezner and Henry Farrell. The power and politics of blogs. *Public Choice*, 134(1–2), 2004. 177
- [70] Patrick Dubroy and Ravin Balakrishnan. A study of tabbed browsing among mozilla firefox users. In *Proc. Conference on Human Factors in Computing Systems, SIGCHI*, pages 673–682. ACM, 2010. 3, 47
- [71] Kevin Duh, Tsutomu Hirao, Akisato Kimura, Katsuhiko Ishiguro, Tomoharu Iwata, and Ching-Man Au Yeung. Creating stories: Social curation of twitter messages. In *Proc. Conference on Weblogs and Social Media, ICWSM. AAAI*, 2012. 206
- [72] Georges Dupret and Mounia Lalmas. Absence time and user engagement: evaluating ranking functions. In *Proc. Conference on Web Search and Data Mining, WSDM*, pages 173–182. ACM, 2013. 16, 56, 76, 198
- [73] Andy Edmonds, Ryen W White, Dan Morris, and Steven M Drucker. Instrumenting the dynamic web. *Journal of Web Engineering*, 6(3): 243, 2007. 16, 42
- [74] Claudia Ehmke and Stephanie Wilson. Identifying web usability problems from eye-tracking data. In *Proc. Conference People and Com-*

- puters, BCS-HCI, pages 119–128. British Computer Society, 2007. 15
- [75] Xiao Fang, Paul Jen-Hwa Hu, Michael Chau, Han-Fen Hu, Zhuo Yang, and Olivia R Liu Sheng. A data-driven approach to measure web site navigability. *Journal of Management Information Systems*, 29(2):173–212, 2012. 18
- [76] Ignacio Fernández-Tobías, Iván Cantador, Marius Kaminskas, and Francesco Ricci. Cross-domain recommender systems: A survey of the state of the art. In *Proc. Spanish Conference on Information Retrieval*, CERI, 2012. 77
- [77] Mark Fishman. *Manufacturing the news*, volume 143. University of Texas Press Austin, 1980. 204
- [78] Gemma Fitzsimmons, Mark Weal, and Denis Drieghe. On measuring the impact of hyperlinks on reading. In *Proc. Web Science Conference*, WebSci, pages 65–74. ACM, 2013. 18
- [79] Gemma Fitzsimmons, Mark J. Weal, and Denis Drieghe. Skim reading: An adaptive strategy for reading on the web. In *Proc. Web Science Conference*, WebSci, pages 211–219. ACM, 2014. 18, 55
- [80] Lucie Flekova, Oliver Ferschke, and Iryna Gurevych. What makes a good biography?: multidimensional quality analysis based on wikipedia article feedback data. In *Proc. Conference on World Wide Web*, WWW, pages 855–866. ACM, 2014. 129
- [81] Tom Foran. Native advertising strategies for mobile devices. <http://www.forbes.com/sites/ciocentral/2013/03/14/native-advertising-strategies-for-mobile-devices/>, 2013. 100, 102
- [82] Steve Fox, Kuldeep Karnawat, Mark Mydland, Susan Dumais, and Thomas White. Evaluating implicit measures to improve web search. *ACM Transactions on Information Systems*, 23(2):147–168, 2005. 17
- [83] Linton C Freeman. Centrality in social networks conceptual clarification. *Social networks*, 1(3):215–239, 1979. 85
- [84] Jill Freyne, Michal Jacovi, Ido Guy, and Werner Geyer. Increasing engagement through early recommender intervention. In *Proc. Conference on Recommender systems*, RecSys, pages 85–92. ACM, 2009. 4, 16, 19, 153
- [85] Jerome H Friedman. Stochastic gradient boosting. *Computational Statistics and Data Analysis*, 38(4):367–378, 2002. 114
- [86] Boudhayan Ganguly, Satya Bhusan Dash, Dianne Cyr, and Milena Head. The effects of website design on purchase intention in online shopping: the mediating role of trust and the moderating role of cul-

- ture. *International Journal of Electronic Business*, 8(4):302–330, 2010. 18
- [87] Andrés García-Silva, Jeon-Hyung Kang, Kristina Lerman, and Oscar Corcho. Characterising emergent semantics in twitter lists. In *The Semantic Web: Research and Applications*, ESWC, pages 530–544. Springer, 2012. 206
- [88] David Gefen, Elena Karahanna, and Detmar W Straub. Trust and tam in online shopping: an integrated model. *MIS quarterly*, 27(1): 51–90, 2003. 17
- [89] Ahmad Ghandour, George L Benwell, and Kenneth R Deans. The relationship between website metrics and the financial performance of online businesses. In *Proc. Conference on Information Systems*, ICIS, 2010. 17
- [90] Saptarshi Ghosh, Naveen Sharma, Fabricio Benevenuto, Niloy Ganguly, and Krishna Gummadi. Cognos: crowdsourcing search for topic experts in microblogs. In *Proc. Conference on Research and Development in Information Retrieval*, SIGIR, pages 575–590. ACM, 2012. 207, 230
- [91] Daniel G Goldstein, R Preston McAfee, and Siddharth Suri. The cost of annoying ads. In *Proc. Conference on World Wide Web*, WWW, pages 459–470. ACM, 2013. 19, 103
- [92] Derek Greene, Derek O’Callaghan, and Pádraig Cunningham. Identifying topical twitter communities via user list aggregation. arXiv preprint arXiv:1207.0017, 2012. 206
- [93] Aditi Gupta, Anupam Joshi, and Ponnurangam Kumaraguru. Identifying and characterizing user communities on twitter during crisis events. In *Proc. Conference on Information and Knowledge Management*, CIKM, pages 23–26. ACM, 2012. 206
- [94] Karl Gyllstrom and Marie-Francine Moens. Surfin’ wikipedia: an analysis of the wikipedia (non-random) surfer’s behavior from aggregate access data. In *Proc. Symposium on Information Interaction in Context Symposium*, IiX, pages 155–163. ACM, 2012. 128, 147
- [95] Aaron Halfaker, Oliver Keyes, and Dario Taraborelli. Making peripheral participation legitimate: reader engagement experiments in wikipedia. In *Proc. Conference on Computer Supported Cooperative Work*, CSCW, pages 849–860. ACM, 2013. 126
- [96] Richard HR Harper. *Being human: Human-computer interaction in the year 2020*. Microsoft Research Limited, 2008. 18

- [97] Angela V Hausman and Jeffrey Sam Siekpe. The effect of web interface features on consumer online purchase intentions. *Journal of Business Research*, 62(1):5–13, 2009. 18
- [98] Brian Haven and Suresh Vittal. Measuring engagement. Forrester Research, 2008. 16
- [99] Denis Helic. Analyzing user click paths in a wikipedia navigation game. In *Proc. International Convention, MIPRO*, pages 374–379. IEEE, 2012. 128
- [100] Eelco Herder. Characterizations of user web revisit behavior. In *Proc. Workshop on Adaptivity and User Modeling in Interactive Systems (WWW)*, ABIS, 2005. 47, 49
- [101] Julia Hoxha. *Cross-domain Recommendations based on semantically-enhanced User Web Behavior*. PhD thesis, Karlsruher Institut für Technologie (KIT), Karlsruhe, 2014. 77
- [102] C Hsieh, Christopher Moghbel, Jianhong Fang, and Junghoo Cho. Experts vs the crowd: Examining popular news prediction performance on twitter. In *Proc. Conference on World Wide Web, WWW*. ACM, 2013. 207
- [103] Po Hu, Minlie Huang, Peng Xu, Weichang Li, Adam K Usadi, and Xiaoyan Zhu. Generating breakpoint-based timeline overview for news topic retrospection. In *Proc. Conference on Data Mining, ICDM*, pages 260–269. IEEE, 2011. 176
- [104] Chun-Yao Huang, Yung-Cheng Shen, I Chiang, Chen-Shun Lin, et al. Concentration of web users' online information behaviour. *Information Research*, 12(4), 2007. 77
- [105] Jeff Huang and Ryen W White. Parallel browsing behavior on the web. In *Proc. Conference on Hypertext and Hypermedia, HT*, pages 13–18. ACM, 2010. 3, 47, 51
- [106] Jeff Huang, Ryen W White, and Susan Dumais. No clicks, no problem: using cursor movements to understand and improve search. In *Proc. Conference on Human Factors in Computing Systems, SIGCHI*, pages 1225–1234. ACM, 2011. 15
- [107] BA Huberman and DM Wilkinson. Assessing the value of cooperation in wikipedia. *First Monday*, 12(4), 2007. 126
- [108] Samuel Ieong, Mohammad Mahdian, and Sergei Vassilvitskii. Advertising in a stream. In *Proc. Conference on World Wide Web, WWW*, pages 29–38. ACM, 2014. 19, 102
- [109] Alicia Iriberry and GONDY Leroy. A life-cycle perspective on online

- community success. *Computing Surveys (CSUR)*, 41(2):11, 2009. 126
- [110] Richard Jacques et al. Engagement as a design concept for multimedia. *Canadian Journal of Educational Communication*, 24(1):49–59, 1995. 14
- [111] Bernard J Jansen, Amanda Spink, and Vinish Kathuria. How to define searching sessions on web search engines. In *Advances in Web Mining and Web Usage Analysis*, pages 92–109. Springer, 2007. 60
- [112] Bernard J. Jansen, Zhe Liu, and Zach Simon. The effect of ad rank on the performance of keyword advertising campaigns. *Journal of the American Society for Information Science and Technology*, 64(10): 2115–2132, 2013. 102
- [113] Akshay Java, Xiaodan Song, Tim Finin, and Belle Tseng. Why we twitter: An analysis of a microblogging community. In *Advances in Web Mining and Web Usage Analysis*, pages 118–138. Springer, 2009. 204
- [114] Qiqi Jiang, Chuan-Hoo Tan, and Kwok-Kee Wei. Cross-website navigation behavior and purchase commitment: A pluralistic field research. In *Proc. Pacific Asia Conference on Information Systems, PACIS*, 2012. 77
- [115] Eric J Johnson, Wendy W Moe, Peter S Fader, Steven Bellman, and Gerald L Lohse. On the depth and dynamics of online search behavior. *Management Science*, 50(3):299–308, 2004. 77
- [116] Aaron A Jones. *The Impact of website navigational usability characteristics on user frustration and performance metrics*. PhD thesis, Ohio University, 2012. 18
- [117] Andrew Kae, Kin Kan, Vijay K Narayanan, and Dragomir Yankov. Categorization of display ads using image and landing page features. In *Proc. Workshop on Large Scale Data Mining: Theory and Applications*, page 1. ACM, 2011. 102, 104, 112, 114
- [118] James Kalbach. *Designing web navigation*. O’Reilly Media, Inc., 2007. 4, 18
- [119] Andreas Kaltenbrunner, Vicenç Gómez, Ayman Moghnieh, Rodrigo Meza, Josep Blat, and Vicente López. Homogeneous temporal activity patterns in a large online communication space. *arXiv preprint arXiv:0708.1579*, 2007. 176
- [120] Maryam Kamvar and Shumeet Baluja. A large scale study of wireless search behavior: Google mobile search. In *Proc. Conference on Human Factors in Computing Systems, SIGCHI*, pages 701–709. ACM, 2006.

102

- [121] Avinash Kaushik. Standard metrics revisited: #4 : Time on page & time on site. <http://www.kaushik.net/avinash/standard-metrics-revisited-time-on-page-and-time-on-site/>, 2008. 22, 52
- [122] Avinash Kaushik. *Web Analytics 2.0: The Art of Online Accountability and Science of Customer Centricity*. SYBEX Inc., 2009. 2, 42, 160
- [123] Brian Keegan, Darren Gergle, and Noshir Contractor. Hot off the wiki: Structures and dynamics of wikipedia’s coverage of breaking news events. *American Behavioral Scientist*, 2013. 139
- [124] Melanie Kellar, Carolyn Watters, and Michael Shepherd. A goal-based classification of web information tasks. *American Society for Information Science and Technology (ASIS&T)*, 43(1):1–22, 2006. 27
- [125] David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *Proc. Conference on Knowledge Discovery and Data Mining, SIGKDD*, pages 137–146. ACM, 2003. 77
- [126] Michael Khoo, Joe Pagano, Anne L Washington, Mimi Recker, Bart Palmer, and Robert A Donahue. Using web metrics to analyze digital libraries. In *Proc. Conference on Digital Libraries*, pages 375–384. ACM/IEEE, 2008. 17, 34
- [127] Dongwoo Kim, Yohan Jo, Il-Chul Moon, and Alice Oh. Analysis of twitter lists as a potential source for discovering latent characteristics of users. In *Proc. Workshop on Microblogging, SIGCHI*. ACM, 2010. 206
- [128] Youngho Kim, Ahmed Hassan, Ryen W White, and Imed Zitouni. Modeling dwell time to predict click-level satisfaction. In *Proc. Conference on Web Search and Data Mining, WSDM*, pages 193–202. ACM, 2014. 103
- [129] Aniket Kittur and Robert E Kraut. Harnessing the wisdom of crowds in wikipedia: quality through coordination. In *Proc. Conference on Computer Supported Cooperative Work, CSCW*, pages 37–46. ACM, 2008. 126
- [130] Ron Kohavi, Alex Deng, Brian Frasca, Roger Longbotham, Toby Walker, and Ya Xu. Trustworthy online controlled experiments: Five puzzling outcomes explained. In *Proc. Conference on Knowledge Discovery and Data Mining, SIGKDD*, pages 786–794. ACM, 2012. 118
- [131] Kevin Koidl, Owen Conlan, and Vincent Wade. Cross-site person-

- alization: assisting users in addressing information needs that span independently hosted websites. In *Proc. Conference on Hypertext and Hypermedia*, HT, pages 66–76. ACM, 2014. 77
- [132] Peter Kollock. The economies of online cooperation: Gifts and public goods in cyberspace. *Communities in Cyberspace*, page 220, 1999. 126
- [133] Shoubin Kong and Ling Feng. A tweet-centric approach for topic-specific author ranking in micro-blog. In *Advanced Data Mining and Applications*, ADMA, pages 138–151. Springer, 2011. 207
- [134] Alok Kothari, Walid Magdy, Kareem Darwish, Ahmed Mourad, and Ahmed Taei. Detecting comments on news articles in microblogs. In *Proc. Conference on Weblogs and Social Media*, ICWSM. AAAI, 2013. 204
- [135] Marios Koufaris and William Hampton-Sosa. The development of initial trust in an online company by new customers. *Information & Management*, 41(3):377–397, 2004. 17
- [136] Steve Krug. *Don't make me think!: a common sense approach to Web usability*. Pearson Education India, 2000. 18
- [137] Anagha Kulkarni, Jaime Teevan, Krysta M Svore, and Susan T Dumais. Understanding temporal query dynamics. In *Proc. Conference on Web Search and Data Mining*, WSDM, pages 167–176. ACM, 2011. 34
- [138] Ravi Kumar and Andrew Tomkins. A characterization of online browsing behavior. In *Proc. Conference on World Wide Web*, WWW, pages 561–570. ACM, 2010. 2, 24, 47, 52, 76
- [139] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In *Proc. Conference on World Wide Web*, WWW, pages 591–600. ACM, 2010. 3, 204, 226
- [140] Mounia Lalmas and Janette Lehmann. Models of user engagement. In Heather L O'Brien and Mounia Lalmas, editors, *Why Engagement Matters: Cross-disciplinary Perspectives and Innovations on User Engagement with Digital Media*. Springer, 2015. (in progress). 5
- [141] Mounia Lalmas, Janette Lehmann, Guy Shaked, Fabrizio Silvestri, and Gabriele Tolomei. Measuring post-click user experience with mobile native advertising on streams. submitted for publication, 2014. 7
- [142] Mounia Lalmas, Heather L O'Brien, and Elad Yom-Tov. *Measuring user engagement*. Synthesis Lectures on Sample Series #1. Morgan and cLaypool publishers, 2014. 14, 16

- [143] Brenda Laurel. *Computers as theatre*. Pearson Education, 2013. 13
- [144] Jean Lave and Etienne Wenger. *Situated learning: Legitimate peripheral participation*. Cambridge University Press, 1991. 126
- [145] Effie Lai-Chong Law, Virpi Roto, Marc Hassenzahl, Arnold POS Vermeeren, and Joke Kort. Understanding, scoping and defining user experience: a survey approach. In *Proc. Conference on Human Factors in Computing Systems*, SIGCHI, pages 719–728. ACM, 2009. 18
- [146] Jung-Hyun Lee, Jongwoo Ha, Jin-Yong Jung, and Sangkeun Lee. Semantic contextual advertising based on the open directory project. *ACM Transactions on the Web (TWEB)*, 7(4):24, 2013. 104
- [147] Sang-Myung Lee, Gerardo R Ungson, and Michael V Russo. What determines an engaging website?: An empirical study of website characteristics and operational performance. *The Journal of High Technology Management Research*, 22(1):67–79, 2011. 18, 153
- [148] Janette Lehmann, Mounia Lalmas, Elad Yom-Tov, and Georges Dupret. Models of user engagement. In *Proc. Conference on User Modeling, Adaptation, and Personalization*, UMAP, pages 164–175. Springer, 2012. 5
- [149] Janette Lehmann, Carlos Castillo, Mounia Lalmas, and Ethan Zuckerman. Transient news crowds in social media. In *Proc. Conference on Weblogs and Social Media*, ICWSM. AAAI, 2013. 9
- [150] Janette Lehmann, Carlos Castillo, Mounia Lalmas, and Ethan Zuckerman. Finding news curators in twitter. In *Proc. Conference on World Wide Web Companion*, WWW Companion, pages 863–870. ACM, 2013. 9
- [151] Janette Lehmann, Mounia Lalmas, and Ricardo Baeza-Yates. Temporal variations in networked user engagement. In *TNETS Satellite at European Conference on Complex Systems (ECCS)*, 2013. 7
- [152] Janette Lehmann, Mounia Lalmas, Ricardo Baeza-Yates, and Elad Yom-Tov. Networked user engagement. In *Proc. Workshop on User engagement optimization at CIKM*, pages 7–10. ACM, 2013. 7
- [153] Janette Lehmann, Mounia Lalmas, Georges Dupret, and Ricardo Baeza-Yates. Online multitasking and user engagement. In *Proc. Conference on Information and Knowledge Management*, CIKM, pages 519–528. ACM, 2013. 6
- [154] Janette Lehmann, Carlos Castillo, Mounia Lalmas, , and Ricardo Baeza-Yates. Story-focused reading in online news. submitted for publication, 2014. 9

- [155] Janette Lehmann, Mounia Lalmas, and Ricardo Baeza-Yates. Measuring inter-site engagement. In V. Govindaraju, V. V. Raghavan, and C. R. Rao, editors, *Handbook of Statistics*. Elsevier, 2014. 7, 8
- [156] Janette Lehmann, Claudia Müller-Birn, David Laniado, Mounia Lalmas, and Andreas Kaltenbrunner. Reader preferences and behavior on wikipedia. In *Proc. Conference on Hypertext and Hypermedia*, HT, pages 88–97. ACM, 2014. 8
- [157] Janette Lehmann, Claudia Müller-Birn, David Laniado, Mounia Lalmas, and Andreas Kaltenbrunner. What and how users read: Transforming reading behavior into valuable feedback for the wikipedia community. Presentation at Wikimania, 2014. 8
- [158] Jure Leskovec, Lars Backstrom, and Jon Kleinberg. Meme-tracking and the dynamics of the news cycle. In *Proc. Conference on Knowledge Discovery and Data Mining*, SIGKDD, pages 497–506. ACM, 2009. 176
- [159] Mark Levene. *An Introduction to Search Engines and Web Navigation*. Addison Wesley, 2005. 105
- [160] Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proc. Conference on World Wide Web*, WWW, pages 661–670. ACM, 2010. 175
- [161] Ting-Peng Liang, Hung-Jen Lai, and Yi-Cheng Ku. Personalized content recommendation and user satisfaction: Theoretical synthesis and empirical findings. *Journal of Management Information Systems*, 23(3):45–70, 2007. 4, 19
- [162] Q Vera Liao, Claudia Wagner, Peter Pirolli, and Wai-Tat Fu. Understanding experts’ and novices’ expertise judgment of twitter users. In *Proc. Conference on Human Factors in Computing Systems*, SIGCHI, pages 2461–2464. ACM, 2012. 207
- [163] Gitte Lindgaard, Gary Fernandes, Cathy Dudek, and Judith Brown. Attention web designers: You have 50 milliseconds to make a good first impression! *Behaviour & information technology*, 25(2):115–126, 2006. 18
- [164] Siân E Lindley, Sam Meek, Abigail Sellen, and Richard Harper. It’s simply integral to what i do: enquiries into how the web is weaved into everyday life. In *Proc. Conference on World Wide Web*, WWW, pages 1067–1076. ACM, 2012. 24, 27
- [165] Haibin Liu, Woo-Cheol Kim, and Dongwon Lee. Characterizing landing pages in sponsored search. In *Proc. Latin American Web Congress*,

- LA-WEB, pages 100–107. IEEE, 2012. 111
- [166] Jiahui Liu, Peter Dolan, and Elin Rønby Pedersen. Personalized news recommendation based on click behavior. In *Proc. Conference on Intelligent User Interfaces*, IUI, pages 31–40. ACM, 2010. 175, 184
- [167] Jingjing Liu and Nicholas J Belkin. Personalizing information retrieval for multi-session tasks: The roles of task stage and task type. In *Proc. Conference on Research and Development in Information Retrieval*, SIGIR, pages 26–33. ACM, 2010. 2, 16
- [168] Yuting Liu, Bin Gao, Tie-Yan Liu, Ying Zhang, Zhiming Ma, Shuyuan He, and Hang Li. Browserank: letting web users vote for page importance. In *Proc. Conference on Research and Development in Information Retrieval*, SIGIR, pages 451–458. ACM, 2008. 77, 162
- [169] Gilad Lotan, Devin Gaffney, and Cherie Meyer. Audience analysis of major news accounts on twitter. *Social Flow*, 2011. 3, 211
- [170] Claudio Lucchese, Salvatore Orlando, Raffaele Perego, Fabrizio Silvestri, and Gabriele Tolomei. Identifying task-based sessions in search engine query logs. In *Proc. Conference on Web Search and Data Mining*, WSDM, pages 277–286. ACM, 2011. 27, 47
- [171] Brian W Matthews. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochim Biophys Acta*, 405 (2):442–451, 1975. 115
- [172] Lori McCay-Peet, Mounia Lalmas, and Vidhya Navalpakkam. On saliency, affect and focused attention. In *Proc. Conference on Human Factors in Computing Systems*, SIGCHI, pages 541–550. ACM, 2012. 4
- [173] Ben McConnell and Jackie Huba. The 1% rule: Charting citizen participation. Church of the Customer Blog, 2006. 205
- [174] Richard McCreadie, Craig Macdonald, and Iadh Ounis. News vertical search: when and what to display to users. In *Proc. Conference on Research and Development in Information Retrieval*, SIGIR, pages 253–262. ACM, 2013. 19, 174, 177
- [175] Bruce McKenzie and Andy Cockburn. An empirical analysis of web page revisitation. In *Proc. Conference on System Sciences*, HICSS. IEEE, 2001. 47
- [176] David Mehrzadi and Dror G Feitelson. On extracting session data from activity logs. In *Proc. Conference on Systems and Storage*, SYSTOR, page 3. ACM, 2012. 52
- [177] Mark Meiss, John Duncan, Bruno Gonçalves, José J Ramasco, and

- Filippo Menczer. What's in a session: tracking individual behavior on the web. In *Proc. Conference on Hypertext and Hypermedia*, HT, pages 173–182. ACM, 2009. 46, 47
- [178] Mark R Meiss, Filippo Menczer, Santo Fortunato, Alessandro Flammini, and Alessandro Vespignani. Ranking web sites with real user traffic. In *Proc. Conference on Web Search and Data Mining*, WSDM, pages 65–76. ACM, 2008. 77
- [179] Mark R Meiss, Bruno Gonçalves, José J Ramasco, Alessandro Flammini, and Filippo Menczer. Agents, bookmarks and clicks: a topical model of web navigation. In *Proc. Conference on Hypertext and Hypermedia*, HT, pages 229–234. ACM, 2010. 50, 51, 76
- [180] Eric Meyerson. Youtube now: Why we focus on watch time. <http://youtubecreator.blogspot.co.uk/2012/08/youtube-now-why-we-focus-on-watch-time.html>, 2012. 16
- [181] Matthew Michelson and Sofus A Macskassy. Discovering users' topics of interest on twitter: a first look. In *Proc. Workshop on Analytics for Noisy Unstructured Text Data*, AND, pages 73–80. ACM, 2010. 206
- [182] David Milne and Ian H Witten. Learning to link with wikipedia. In *Proc. Conference on Information and Knowledge Management*, CIKM, pages 509–518. ACM, 2008. 177
- [183] Sunil Mithas, Narayan Ramasubbu, Mayuram S Krishnan, and Claes Fornell. Designing web sites for customer loyalty across business domains: a multilevel analysis. *Journal of Management Information Systems*, 23(3):97–127, 2007. 18
- [184] Wendy W Moe and Peter S Fader. Capturing evolving visit behavior in clickstream data. *Journal of Interactive Marketing*, 18(1):5–19, 2004. 17
- [185] Sai T Moturu and Huan Liu. Quantifying the trustworthiness of social media content. *Distributed and Parallel Databases*, 29(3):239–260, 2011. 19
- [186] Vanessa Murdock, Massimiliano Ciaramita, and Vassilis Plachouras. A noisy-channel approach to contextual advertising. In *Proc. Workshop on Data mining and audience intelligence for advertising*, pages 21–27. ACM, 2007. 104
- [187] Vidhya Navalpakkam and Elizabeth Churchill. Mouse tracking: measuring and predicting users' experience of web-based content. In *Proc. Conference on Human Factors in Computing Systems*, SIGCHI, pages 2963–2972. ACM, 2012. 15

- [188] Vidhya Navalpakkam, LaDawn Jentzsch, Rory Sayres, Sujith Ravi, Amr Ahmed, and Alex Smola. Measurement and modeling of eye-mouse behavior in the presence of nonlinear page layouts. In *Proc. Conference on World Wide Web, WWW*, pages 953–964. ACM, 2013. 15
- [189] Oriella PR Network. The influence game: How news is sources and managed today. <http://www.oriellaprnnetwork.com/sites/default/files/research/OriellaDigitalJournalismStudy2012FinalUS.pdf>, 2012. 203
- [190] Mark EJ Newman. The structure and function of complex networks. *SIAM review*, 45(2):167–256, 2003. 76, 77, 165
- [191] Zezia Benhamza Nsairi and Manel Khadraoui. Website satisfaction: Determinants and consequences on website loyalty. *International Business Research*, 6(9):p77, 2013. 18
- [192] Hartmut Obendorf, Harald Weinreich, Eelco Herder, and Matthias Mayer. Web page revisitation revisited: implications of a long-term click-stream study of browser usage. In *Proc. Conference on Human Factors in Computing Systems, SIGCHI*, pages 597–606. ACM, 2007. 3, 46, 47, 51, 54
- [193] Heather L O’Brien. The influence of hedonic and utilitarian motivations on user engagement: The case of online shopping experiences. *Interacting with Computers*, 22(5):344–352, 2010. 15, 153
- [194] Heather L O’Brien. Exploring user engagement in online news interactions. *American Society for Information Science and Technology (ASIS&T)*, 48(1):1–10, 2011. 175
- [195] Heather L O’Brien and Elaine G Toms. What is user engagement? a conceptual framework for defining user engagement with technology. *American Society for Information Science and Technology (ASIS&T)*, 59(6):938–955, 2008. 1, 14
- [196] Heather L O’Brien and Elaine G Toms. The development and evaluation of a survey to measure user engagement. *American Society for Information Science and Technology (ASIS&T)*, 61(1):50–69, 2010. 15
- [197] Heather L O’Brien and Elaine G Toms. Examining the generalizability of the user engagement scale (ues) in exploratory search. *Information Processing & Management*, 49(5):1092–1107, 2013. 15, 18
- [198] Richard J Oentaryo, Ee-Peng Lim, Jia-Wei Low, David Lo, and Michael Finegold. Predicting response in mobile advertising with hierarchical importance-aware factorization machine. In *Proc. Conference*

- on *Web Search and Data Mining*, WSDM, pages 123–132. ACM, 2014. 102
- [199] Chitu Okoli, Mohamad Mehdi, Mostafa Mesgari, Finn Nielsen, and Arto Lanamäki. The people’s encyclopedia under the gaze of the sages: A systematic review of scholarly research on wikipedia. *Social Science Research Network (SSRN)*, 2012. 126
- [200] Nuria Oliver, Greg Smith, Chintan Thakkar, and Arun C Surendran. Swish: semantic analysis of window titles and switching history. In *Proceedings of the 11th international conference on Intelligent user interfaces*, pages 194–201. ACM, 2006. 45
- [201] Michaël Opgenhaffen and Leen d’Haenens. The impact of online news features on learning from news: a knowledge experiment. *Journal of Internet Science*, 6(1):8–28, 2011. 176
- [202] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab, 1999. 18, 77, 86
- [203] Aditya Pal and Scott Counts. Identifying topical authorities in microblogs. In *Proc. Conference on Web Search and Data Mining*, WSDM, pages 45–54. ACM, 2011. 207
- [204] Young-Hoon Park and Peter S Fader. Modeling browsing behavior at multiple websites. *Marketing Science*, 23(3):280–303, 2004. 77
- [205] Alex Penev and Raymond K Wong. Framework for timely and accurate ads on mobile devices. In *Proc. Conference on Knowledge Discovery and Data Mining*, SIGKDD, pages 1067–1076. ACM, 2009. 102
- [206] Eric T Peterson and Joseph Carrabis. Measuring the immeasurable: Visitor engagement. *Web Analytics Demystified*, 2008. 14, 16
- [207] Owen Phelan, Kevin McCarthy, Mike Bennett, and Barry Smyth. Terms of a feather: Content-based news recommendation and discovery using twitter. In *Advances in Information Retrieval*, pages 448–459. Springer, 2011. 206
- [208] Joshua Porter. *Designing for the social web*. Peachpit Press, 2010. 17, 18, 19, 68
- [209] Poynter.org. Survey: Americans turn to established media for breaking news, mobile. <http://www.poynter.org/latest-news/top-stories/190586/new-data-show-shifting-patterns-as-people-look-for-news-across-platforms/>, 2012. 173
- [210] Jennifer Preece and Ben Shneiderman. The reader-to-leader framework: Motivating technology-mediated social participation. *Trans-*

- actions on Human-Computer Interaction (TOCHI)*, 1(1):13–32, 2009. 126, 149
- [211] Jenny Preece, Blair Nonnecke, and Dorine Andrews. The top five reasons for lurking: improving community experiences for everyone. *Computers in Human Behavior*, 20(2):201–223, 2004. 126
- [212] Hemant Purohit, Yiye Ruan, Amruta Joshi, Srinivasan Parthasarathy, and Amit Sheth. Understanding user-community engagement by multi-faceted features: A case study on twitter. In *Proc. Workshop on Social Media Engagement (WWW)*, SoME, 2011. 206
- [213] Filip Radlinski and Thorsten Joachims. Query chains: learning to rank from implicit feedback. In *Proc. Conference on Knowledge Discovery and Data Mining, SIGKDD*, pages 239–248. ACM, 2005. 176
- [214] Jacob Ratkiewicz, Santo Fortunato, Alessandro Flammini, Filippo Menczer, and Alessandro Vespignani. Characterizing and modeling the dynamics of online popularity. *Physical Review Letters*, 105(15):158701, 2010. 128, 133, 141
- [215] JC Read, SJ MacFarlane, and Chris Casey. Endurability, engagement and expectations: Measuring children’s fun. In *Interaction design and children*, volume 2, pages 1–23. Shaker Publishing Eindhoven, 2002. 13, 16
- [216] Antonio José Reinoso Peinado. *Temporal and behavioral patterns in the use of Wikipedia*. PhD thesis, Universidad Rey Juan Carlos, 2011. 128, 139
- [217] Matthew Richardson, Ewa Dominowska, and Robert Ragno. Predicting clicks: estimating the click-through rate for new ads. In *Proc. Conference on World Wide Web, WWW*, pages 521–530. ACM, 2007. 104
- [218] Lukas Ritzel, Cem Van der Schaar, and Steven Goodman. Native advertising mobil. Hochschule für Wirtschaft Zürich Zürich, Schweiz, 2013. 102
- [219] Kerry Rodden, Hilary Hutchinson, and Xin Fu. Measuring the user experience on a large scale: user-centered metrics for web applications. In *Proc. Conference on Human Factors in Computing Systems, CHI*, pages 2395–2398. ACM, 2010. 2, 13
- [220] Jason MT Roos. *Hyper-Media Search and Consumption*. PhD thesis, Duke University, 2012. 77, 177
- [221] Rómer Rosales, Haibin Cheng, and Eren Manavoglu. Post-click conversion modeling and analysis for non-guaranteed delivery display

- advertising. In *Proc. Conference on Web Search and Data Mining, WSDM*, pages 293–302. ACM, 2012. 102, 103
- [222] Marco MC Rozendaal, David V Keyson, and Huib de Ridder. Product features and task effects on experienced richness, control and engagement in voicemail browsing. *Personal and Ubiquitous Computing*, 13(5):343–354, 2009. 14
- [223] Joshua S Rubinstein, David E Meyer, and Jeffrey E Evans. Executive control of cognitive processes in task switching. *Journal of Experimental Psychology: Human Perception and Performance*, 27(4):763, 2001. 45
- [224] Sivan Sabato, Elad Yom-Tov, Aviad Tsherniak, and Saharon Rosset. Analyzing system logs: A new view of what’s important. In *Proc. Workshop on Tackling computer systems problems with machine learning techniques*, USENIX, pages 1–7. USENIX Association, 2007. 36
- [225] Diego Saez-Trumper, Carlos Castillo, and Mounia Lalmas. Social media news communities: gatekeeping, coverage, and statement bias. In *Proc. Conference on Information and Knowledge Management, CIKM*, pages 1679–1684. ACM, 2013. 173
- [226] David Sasaki. Our friends become curators of Twitter-based news. <http://www.pbs.org/idealab/2010/04/our-friends-become-curators-of-twitter-based-news092.html>, April 2010. 226
- [227] Yuki Sato, Daisuke Yokomoto, Hiroyuki Nakasaki, Mariko Kawaba, Takehito Utsuro, and Tomohiro Fukuhara. Linking topics of news and blogs with wikipedia for complementary navigation. In *Social Software: Recent Trends and Developments in Social Software*, pages 75–87. Springer, 2011. 177
- [228] Jeff Sauro and Joseph S Dumas. Comparison of three one-question, post-task usability questionnaires. In *Proc. Conference on Human Factors in Computing Systems, SIGCHI*, pages 1599–1608. ACM, 2009. 15
- [229] D Sculley, Robert G Malkin, Sugato Basu, and Roberto J Bayardo. Predicting bounce rates in sponsored search advertisements. In *Proc. Conference on Knowledge Discovery and Data Mining, SIGKDD*, pages 1325–1334. ACM, 2009. 16, 103, 104
- [230] Dafna Shahaf and Carlos Guestrin. Connecting the dots between news articles. In *Proc. Conference on Knowledge Discovery and Data Mining, SIGKDD*, pages 623–632. ACM, 2010. 176

- [231] Naveen Kumar Sharma, Saptarshi Ghosh, Fabricio Benevenuto, Niloy Ganguly, and Krishna Gummadi. Inferring who-is-who in the twitter social network. *Computer Communication Review (SIGCOMM)*, 42(4):533–538, 2012. 229, 231
- [232] MV Simkin and VP Roychowdhury. A theory of web traffic. *Europhysics Letters (EPL)*, 82(2):28006, 2008. 76
- [233] Mel Slater. Measuring presence: A response to the witmer and singer presence questionnaire. *Presence: Teleoperators and Virtual Environments*, 8(5):560–565, 1999. 15
- [234] Michael D Smith and Erik Brynjolfsson. Consumer decision-making at an internet shopbot: Brand still matters. *The Journal of Industrial Economics*, 49(4):541–558, 2001. 17
- [235] Eric Sodomka, Sébastien Lahaie, and Dustin Hillard. A predictive model for advertiser value-per-click in sponsored search. In *Proc. Conference on Information and Knowledge Management, CIKM*, pages 1179–1190. ACM, 2013. 19, 102
- [236] Yang Song, Hao Ma, Hongning Wang, and Kuansan Wang. Exploring and exploiting user search behavior on mobile and tablet devices to improve search relevance. In *Proc. Conference on World Wide Web, WWW*, pages 1201–1212. ACM, 2013. 101, 102, 103, 109, 122
- [237] Amanda Spink, Minsoo Park, Bernard J Jansen, and Jan Pedersen. Multitasking web search on alta vista. In *Proc. Conference on Information Technology: Coding and Computing, ITCC*, pages 309–313. IEEE, 2004. 47
- [238] Anselm Spoerri. Visualizing the overlap between the 100 most visited pages on wikipedia for september 2006 to january 2007. *First Monday*, 12(4), 2007. 127
- [239] Anselm Spoerri. What is popular on wikipedia and why? *First Monday*, 12(4), 2007. 131, 146
- [240] Tiziano Squartini, Francesco Picciolo, Franco Ruzzenenti, and Diego Garlaschelli. Reciprocity of weighted networks. *Nature: Scientific reports*, 3, 2013. 85
- [241] Tadej Štajner, Bart Thomee, Ana-Maria Popescu, Marco Pennacchiotti, and Alejandro Jaimes. Automatic selection of social media responses to news. In *Proc. Conference on Knowledge Discovery and Data Mining, SIGKDD*, pages 50–58. ACM, 2013. 204
- [242] Besiki Stvilia, Michael B Twidale, Linda C Smith, and Les Gasser. Information quality work organization in wikipedia. *American Society*

- for *Information Science and Technology (ASIS&T)*, 59(6):983–1001, 2008. 126
- [243] Ilija Subašić and Bettina Berendt. Peddling or creating? investigating the role of twitter in news reporting. In *Advances in Information Retrieval*, pages 207–213. Springer, 2011. 203
- [244] Bongwon Suh, Ed H Chi, Aniket Kittur, and Bryan A Pendleton. Lifting the veil: improving accountability and social transparency in wikipedia with wikidashboard. In *Proc. Conference on Human Factors in Computing Systems*, SIGCHI, pages 1037–1040. ACM, 2008. 146
- [245] Alistair Sutcliffe. Designing for user engagement: Aesthetic and attractive user interfaces. *Synthesis lectures on human-centered informatics*, 2(1):1–55, 2009. 18
- [246] Chadwyn Tann and Mark Sanderson. Are web-based informational queries changing? *American Society for Information Science and Technology (ASIS&T)*, 60(6):1290–1293, 2009. 128
- [247] Linda Tauscher and Saul Greenberg. How people revisit web pages: Empirical findings and implications for the design of history systems. *Journal of Human-Computer Studies*, 47(1):97–137, 1997. 54
- [248] Marijn ten Thij, Andreas Kaltenbrunner, David Laniado, and Yana Volkovich. Modeling and predicting page-view dynamics on wikipedia. *Computing Research Repository (CoRR)*, abs/1212.5943, 2012. 18, 34, 128, 133
- [249] Narongsak Thongpapanl and Abdul Rehman Ashraf. Enhancing on-line performance through website content and personalization. *Journal of Computer Information Systems*, 52(1):3, 2011. 4, 19
- [250] Noam Tractinsky, AS Katz, and Dror Ikar. What is beautiful is usable. *Interacting with computers*, 13(2):127–145, 2000. 15, 18
- [251] Michele Trevisiol, Luca Chiarandini, Luca Maria Aiello, and Alejandro Jaimes. Image ranking based on user browsing behavior. In *Proc. Conference on Research and Development in Information Retrieval*, SIGIR, pages 445–454. ACM, 2012. 77, 153, 160
- [252] Michele Trevisiol, Luca Maria Aiello, Rossano Schifanella, and Alejandro Jaimes. Cold-start news recommendation with domain-dependent browse graph. In *Proc. Conference on Recommender systems*, RecSys, pages 81–88. ACM, 2014. 19, 77
- [253] Manos Tsagkias, Maarten de Rijke, and Wouter Weerkamp. Linking online news and social media. In *Proc. Conference on Web Search and Data Mining*, WSDM, pages 565–574. ACM, 2011. 204

- [254] Joseph Turow and Lokman Tsui. *The Hyperlinked Society: Questioning Connections in the Digital Age*. The University of Michigan Press, 2008. 18, 154
- [255] Joseph Turow and Lokman Tsui. The hyperlinked society. *The University of Michigan Press*, 2008. 18, 177
- [256] Sarah K Tyler and Jaime Teevan. Large scale query log analysis of re-finding. In *Proc. Conference on Web Search and Data Mining, WSDM*, pages 191–200. ACM, 2010. 47
- [257] Maximilian Viermetz, Carsten Stolz, Vassil Gedov, and Michal Skubacz. Relevance and impact of tabbed browsing behavior on web usage mining. In *Proc. Conference on Web Intelligence*, volume 2006 of *WI. IEEE/WIC/ACM*, 2006. 50
- [258] Christian von der Weth and Manfred Hauswirth. Analysing parallel and passive web browsing behavior and its effects on website metrics. *arXiv preprint arXiv:1402.5255*, 2014. 47
- [259] Claudia Wagner, Vera Liao, Peter Pirolli, Les Nelson, and Markus Strohmaier. It’s not in their tweets: Modeling topical expertise of twitter users. In *Proc. Conference on Social Computing, SocialCom*, pages 91–100. IEEE, 2012. 207, 230
- [260] Vivienne Waller. The search queries that took australian internet users to wikipedia. *Information Research*, 16(2), 2011. 127, 131, 146
- [261] Alex Hai Wang. Don’t follow me: Spam detection in twitter. In *Proc. Conference on Security and Cryptography, SECRCRYPT*, pages 1–10. IEEE, 2010. 209
- [262] Canhui Wang, Min Zhang, Shaoping Ma, and Liyun Ru. Automatic online news issue construction in web environment. In *Proc. Conference on World Wide Web, WWW*, pages 457–466. ACM, 2008. 176
- [263] Jian Wang and Yi Zhang. Utilizing marginal net utility for recommendation in e-commerce. In *Proc. Conference on Research and Development in Information Retrieval, SIGIR*, pages 1003–1012. ACM, 2011. 19
- [264] Qing Wang and Huiyou Chang. Multitasking bar: prototype and evaluation of introducing the task concept into a browser. In *Proc. Conference on Human Factors in Computing Systems, SIGCHI*, pages 103–112. ACM, 2010. 2, 47
- [265] Yang Wang and Alfred Kobsa. Privacy in cross-system personalization. In *Proc. AAAI Spring Symposium: Intelligent Information Privacy Management*, 2010. 77

- [266] Stanley Wasserman. *Social network analysis: Methods and applications*, volume 8. Cambridge University Press, 1994. 84
- [267] Jane Webster and Jaspreet S Ahuja. Enhancing the design of web navigation systems: the influence of user disorientation on engagement and performance. *MIS Quarterly*, pages 661–678, 2006. 4, 18
- [268] Jane Webster and Hayes Ho. Audience engagement in multimedia presentations. *ACM SIGMIS Database*, 28(2):63–77, 1997. 13
- [269] Birgit Weischedel and Eelko KRE Huizingh. Website optimization with web metrics: a case study. In *Proc. Conference on Electronic commerce: The new e-commerce: innovations for conquering current barriers, obstacles and limitations to conducting successful business on the internet*, pages 463–470. ACM, 2006. 16, 22
- [270] Michael J Welch, Uri Schonfeld, Dan He, and Junghoo Cho. Topical semantics of twitter links. In *Proc. Conference on Web Search and Data Mining*, WSDM, pages 327–336. ACM, 2011. 206
- [271] Jianshu Weng, Ee-Peng Lim, Jing Jiang, and Qi He. Twiterrank: finding topic-sensitive influential twitterers. In *Proc. Conference on Web Search and Data Mining*, WSDM, pages 261–270. ACM, 2010. 207, 230
- [272] Robert West, Ingmar Weber, and Carlos Castillo. Drawing a data-driven portrait of wikipedia editors. In *Proc. Symposium on Wikis and Open Collaboration*, Wikisym, page 3. ACM, 2012. 128
- [273] Thomas Wöhner and Ralf Peters. Assessing the quality of wikipedia articles with lifecycle based metrics. In *Proc. Symposium on Wikis and Open Collaboration*, Wikisym, page 16. ACM, 2009. 135
- [274] Shaomei Wu, Jake M Hofman, Winter A Mason, and Duncan J Watts. Who says what to whom on twitter. In *Proc. Conference on World Wide Web*, WWW, pages 705–714. ACM, 2011. 235
- [275] Xiaofei Wu, Ke Yu, and Xin Wang. On the growth of internet application flows: a complex network perspective. In *Proc. Conference on Computer Communications*, INFOCOM, pages 2096–2104. IEEE, 2011. 77
- [276] Elizabeth Yakel. Digital curation. OCLC Systems and Services, 2007. 225
- [277] Xing Yi, Liangjie Hong, Erheng Zhong, Nanthan Nan Liu, and Suju Rajan. Beyond clicks: dwell time for personalization. In *Proc. Conference on Recommender systems*, RecSys, pages 113–120. ACM, 2014. 4, 16, 19, 101, 102, 103

- [278] Peifeng Yin, Ping Luo, Wang-Chien Lee, and Min Wang. Silence is also evidence: interpreting dwell time for recommendation from psychological perspective. In *Proc. Conference on Knowledge Discovery and Data Mining, SIGKDD*, pages 989–997. ACM, 2013. 16, 103
- [279] Elad Yom-Tov, Mounia Lalmas, Georges Dupret, Ricardo Baeza-Yates, Pinard Donmez, and Janette Lehmann. The effect of links on networked user engagement. In *Proc. Conference on World Wide Web (Poster), WWW*, pages 641–642. ACM, 2012. 4, 17, 18, 154
- [280] Elad Yom-Tov, Mounia Lalmas, Ricardo Baeza-Yates, Georges Dupret, Janette Lehmann, and Pinar Donmez. Measuring inter-site engagement. In *Proc. Conference on Big Data, BigData*, pages 228–236. IEEE, 2013. 77, 86, 167, 170
- [281] Hsiang-Fu Yu, Fang-Lan Huang, and Chih-Jen Lin. Dual coordinate descent methods for logistic regression and maximum entropy models. *Machine Learning*, 85(1-2):41–75, 2011. 114
- [282] Barbie Zelizer. Journalists as interpretive communities. *Critical Studies in Media Communication*, 10(3):219–237, 1993. 205
- [283] Haimo Zhang and Shengdong Zhao. Measuring web page revisitation in tabbed browsing. In *Proc. Conference on Human Factors in Computing Systems, SIGCHI*, pages 1831–1834. ACM, 2011. 46, 49
- [284] Yang Zhang, Y Wu, and Q Yang. Community discovery in twitter based on user interests. *Journal of Computational Information Systems*, 8(3):991–1000, 2012. 206
- [285] Wayne Xin Zhao, Jing Jiang, Jianshu Weng, Jing He, Ee-Peng Lim, Hongfei Yan, and Xiaoming Li. Comparing twitter and traditional media using topic models. In *Advances in Information Retrieval*, pages 338–349. Springer, 2011. 204
- [286] Tao Zhu, Patrick Harrington, Junjun Li, and Lei Tang. Bundle recommendation in ecommerce. In *Proc. Conference on Research and Development in Information Retrieval, SIGIR*, pages 657–666. ACM, 2014. 19

Appendix

A.1 Predicting High Quality Ads

The appendix provides detailed information about the ad landing page features, and offline models evaluation of the prediction task of Chapter 7.

A.1.1 Ad Landing Page Features

The prediction model considers the following three types of features:

CONT (C) Features. This group is designed to capture the content of the ad landing page:

- *media*: boolean value stating if the site is responsive.¹
- *clickToCall*: number of clickables linking to a phone call.
- *imageHeight*: height of the landing page.
- *imageWidth*: width of the landing page.
- *numClickable*: number of clickables.
- *numDropdown*: number of dropdown lists.
- *numImages*: number of images.
- *numInputCheckbox*: number of checkboxes.
- *numInputRadio*: number of radio buttons.
- *numInputString*: number of input strings (usually to elicit users details).

¹http://en.wikipedia.org/wiki/Responsive_web_design

- *tokenCount*: number of tokens (words).
- *viewPort*: a boolean feature represents if the site can be tuned to different screen sizes.
- *windowSize*: total width of all `div` tags on the page, which allows detecting carousels.
- *nounsSumOfScores*: number of nouns.
- *numConceptAnnotation*: number of all Wikipedia *entities* (a concept with a Wikipedia entry).
- *summarizabilityScore*: predicts if the page is a good candidate for extracting a summary, where higher value means the landing page is more “newsy”.
- *isMobileOptimised*: the result of the classifier as discussed in Section 7.5.

SIM (S) Features. This group of features captures the similarity of the landing page with the creative text displayed within the stream. Usually, a user sees the creative and decides to click on the basis of the text written there. If the semantics of the creative text is very different from the semantics of the landing page then the user who clicked may be annoyed and leave immediately the page.

- *similarityNoun*: cosine similarity between creative text and landing page based on nouns.
- *similarityWikiIds*: cosine similarity between creative text and landing page based on Wikipedia entities.

HIST (H) Features. These features captures historical information about the ad past performance.

- *impressions*: number of times the ad was shown.
- *clicks*: number of times the ad was clicked.
- *bouncerate*: bounce rate of the ad.
- *avgdwelltime*: average dwell time of the ad.
- *avgdwelltimenonshort*: average dwell time when short clicks were removed.
- *ctr*: click-through rate of the ad.
- *cpx*: cost per click of the ad.

A.1.2 Offline Models Evaluation

The following tables report the results for predicting the probability of high dwell time, low bounce rate, and the combination of them using various configurations.

Table A.1: Dwell Time prediction performance on models built on ads data from March 2014 and tested on April 2014. We vary t_δ to evaluate the impact of the threshold chosen on the prediction ability of the model (best results in bold).

Features	Method	t_δ	AUC	F ₁	MCC
C	logistic	35	0.71	0.65	0.44
C-S	logistic	35	0.70	0.64	0.42
C-H	logistic	35	0.82	0.81	0.64
C-S-H	logistic	35	0.84	0.83	0.67
C	svm	35	0.82	0.81	0.65
C-S	svm	35	0.82	0.81	0.64
C-H	svm	35	0.83	0.82	0.66
C-S-H	svm	35	0.83	0.82	0.66
C	gbdt	35	0.77	0.73	0.55
C-S	gbdt	35	0.77	0.74	0.56
C-H	gbdt	35	0.83	0.82	0.66
C-S-H	gbdt	35	0.83	0.82	0.66
C	logistic	40	0.70	0.60	0.47
C-S	logistic	40	0.72	0.63	0.49
C-H	logistic	40	0.83	0.79	0.66
C-S-H	logistic	40	0.83	0.79	0.66
C	svm	40	0.83	0.79	0.67
C-S	svm	40	0.83	0.79	0.67
C-H	svm	40	0.83	0.80	0.68
C-S-H	svm	40	0.83	0.80	0.68
C	gbdt	40	0.82	0.77	0.70
C-S	gbdt	40	0.81	0.77	0.68
C-H	gbdt	40	0.83	0.80	0.68
C-S-H	gbdt	40	0.83	0.80	0.68
C	logistic	45	0.69	0.57	0.48
C-S	logistic	45	0.70	0.57	0.50
C-H	logistic	45	0.79	0.72	0.60
C-S-H	logistic	45	0.79	0.72	0.60
C	svm	45	0.82	0.76	0.67
C-S	svm	45	0.82	0.76	0.67
C-H	svm	45	0.80	0.73	0.61
C-S-H	svm	45	0.80	0.73	0.61
C	gbdt	45	0.72	0.62	0.56
C-S	gbdt	45	0.71	0.60	0.54
C-H	gbdt	45	0.80	0.73	0.61
C-S-H	gbdt	45	0.80	0.73	0.61

Table A.2: Bounce rate prediction performance on models built on ads data from March 2014 and tested on April 2014. We vary τ_β to evaluate the impact of the threshold chosen on the prediction ability of the model (best result in bold).

Features	Method	τ_β	AUC	F ₁	MCC
C	logistic	0.2	0.51	0.78	0.06
C-S	logistic	0.2	0.59	0.8	0.24
C-H	logistic	0.2	0.79	0.85	0.58
C-S-H	logistic	0.2	0.78	0.85	0.57
C	logistic	0.22	0.61	0.74	0.27
C-S	logistic	0.22	0.61	0.72	0.23
C-H	logistic	0.22	0.86	0.86	0.71
C-R-H	logistic	0.22	0.85	0.86	0.70
C	logistic	0.25	0.57	0.63	0.15
C-S	logistic	0.25	0.63	0.63	0.26
C-H	logistic	0.25	0.83	0.82	0.67
C-S-H	logistic	0.25	0.83	0.81	0.66

Table A.3: High quality (low bounce rate and high dwell time) prediction performance on models built on ads data from March 2014 and tested on April 2014. We fix t_δ to a high value of 50 seconds and we test two τ_β values, 0.1 and 0.2, to evaluate the impact of the threshold chosen on the prediction ability of the model (best results in bold).

Features	Method	t_δ	τ_β	AUC	F ₁	MCC
C	logistic	50	0.1	0.50	0.00	0.06
C-S	logistic	50	0.1	0.50	0.00	0.24
C-H	logistic	50	0.1	0.70	0.43	0.58
C-S-H	logistic	50	0.1	0.63	0.31	0.57
C	svm	50	0.1	0.71	0.50	0.44
C-S	svm	50	0.1	0.71	0.50	0.44
C-H	svm	50	0.1	0.71	0.50	0.59
C-S-H	svm	50	0.1	0.71	0.50	0.59
C	logistic	50	0.2	0.51	0.06	0.06
C-S	logistic	50	0.2	0.56	0.23	0.26
C-H	logistic	50	0.2	0.74	0.58	0.51
C-S-H	logistic	50	0.2	0.74	0.58	0.53
C	svm	50	0.2	0.79	0.67	0.61
C-S	svm	50	0.2	0.79	0.67	0.61
C-H	svm	50	0.2	0.81	0.68	0.62
C-R-H	svm	50	0.2	0.81	0.68	0.62

A.2 Story-focused Reading

The appendix provides the list of news providers used in Chapter 10.

Table A.4: News providers under consideration, listed in alphabetical order.

abcnews.go.com	dallasnews.com	news.sky.com	theguardian.com
adweek.com	denverpost.com	news.yahoo.com	thehill.com
ajc.com	digitalspy.co.uk	newsmax.com	theonion.com
azcentral.com	economist.com	nj.com	thestar.com
bankrate.com	examiner.com	nypost.com	thesundaytimes.co.uk
bbc.co.uk	forbes.com	nytimes.com	time.com
bloomberg.com	foxnews.com	online.wsj.com	upi.com
breitbart.com	heraldsun.com.au	philly.com	usatoday.com
businessweek.com	hollywoodreporter.com	rawstory.com	usnews.com
cbc.ca	huffingtonpost.com	reuters.com	variety.com
cbsnews.com	latimes.com	seattletimes.com	voanews.com
chicagotribune.com	metro.co.uk	sfgate.com	washingtonpost.com
chron.com	miamiherald.com	smh.com.au	washingtontimes.com
cnbc.com	nationalpost.com	theage.com.au	wnd.com
cnn.com	nationalreview.com	theatlantic.com	
csmonitor.com	nbcnews.com	theaustralian.com.au	
dailyfinance.com	news.com.au	theglobeandmail.com	

A.3 Crowd-based News and Curator Discovery

This appendix contains the detailed instructions used to create the training data for the crowd-based news discovery and for the detection of story curators (Chapter 11).

A.3.1 Labelling News Articles

You will be presented with two Twitter messages ("tweets") on current news stories. Please indicate how these two stories are related:

- *Strongly related: Same ongoing news story (e.g. two articles about nuclear inspections in Iran).*
- *Weakly related: Not same story, but same location, person, or topic (e.g. two articles about nuclear proliferation).*
- *Not related.*

Having "Al Jazeera", "BBC", etc. in both tweets does NOT automatically mean they are related.

A.3.2 Labelling News Story Curators

You will be presented with the title of a news article, and tweets and profile information of a Twitter user.

Q1) Please indicate whether the user is interested or an expert of the topic of the article story:

- *Yes: Most of her/his tweets relate to the topic of the story (e.g. the article is about the conflict in Syria, she/he is often tweeting about the conflict in Syria).*
- *Maybe: Many of her/his tweets relate to the topic of the story or she/he is interested in a related topic (e.g. the article is about the conflict in Syria, she/he is tweeting about armed conflicts or the Arabic world).*
- *No: She/he is not tweeting about the topic of the story.*
- *Unknown: Based on the information of the user it was not possible to label her/him.*

Q2) Please indicate whether the user is a human or generates tweets automatically:

- *Human: The user has conversations and personal comments in his tweets. The text of tweets that have URLs (e.g. to news articles) can be self-written and contain own opinions.*
- *Maybe automatic: The Twitter user has characteristics of an automatic profile, but she/he could be human as well.*
- *Automatic: The tweet stream of the user looks automatically generated. The tweets contain only headlines and URLs of news articles.*
- *Unknown: Based on the information of the user it was not possible to label her/him as human or automatic.*

The label “unknown” corresponds to the case where the annotators were not able to reach a decision. Possible reasons were the language of the tweets (e.g. the user is tweeting in Chinese).