Thèse de Doctorat
UPMC Sorbonne Universités - Paris VI

Spécialité

Systèmes Informatiques

présentée par

Alexandru-Florin TATAR

pour obtenir le grade de

Docteur de l'Université Pierre et Marie Curie

Prédiction du Comportement des Utilisateurs: Mobilité et Popularité des Contenus

Soutenue le 9 Juillet 2014 devant le jury composé de :

Walid DABBOUS	**Rapporteur**	**Chercheur, INRIA Sophia Antipolis**
Andrea PASSARELLA	**Rapporteur**	**Chercheur, IIT-CNR**
Martin MAY	**Examinateur**	**Directeur de Stratégie, Technicolor**
Anne-Marie KERMARREC	**Examinateur**	**Directeur de Recherche, INRIA Rennes**
Sébastien TIXEUIL	**Examinateur**	**Professeur, UPMC Sorbonne Universités**
Marcelo DIAS DE AMORIM	**Directeur**	**Directeur de Recherche, UPMC Sorbonne Universités**
Serge FDIDA	**Directeur**	**Professeur, UPMC Sorbonne Universités**

Doctor of Science Thesis

UPMC Sorbonne Universités - Paris VI

Specialization

Computer Science

presented by

Alexandru-Florin Tatar

submitted for the qualification of

Doctor of Science UPMC Sorbonne Universités

Predicting User-Centric Behavior: Mobility and Content Popularity

Committee in charge:

Walid DABBOUS	**Reviewer**	**Senior Researcher, INRIA Sophia Antipolis**
Andrea PASSARELLA	**Reviewer**	**Researcher, IIT-CNR**
Martin MAY	**Examiner**	**Director of Technology Strategy, Technicolor**
Anne-Marie KERMARREC	**Examiner**	**Research Director, INRIA Rennes**
Sébastien TIXEUIL	**Examiner**	**Professor, UPMC Sorbonne Universités**
Marcelo DIAS DE AMORIM	**Advisor**	**Research Director, CNRS and UPMC Sorbonne Universités**
Serge FDIDA	**Advisor**	**Professor, UPMC Sorbonne Universités**

Acknowledgements

This thesis presents the results of my four years of work. It was a long and complex journey; but a very exciting one. Many people have contributed to this work, directly and indirectly, and I am truly grateful to all of them.

In particular, I would like to express my deepest gratitude to my co-advisors, Marcelo Dias de Amorim and Serge Fdida, for accepting me as a PhD student and for their continuous guidance throughout these years. You gave me the freedom to pursue my own ideas and supported me all away this journey. I am not exaggerating when I say that this work would not have been possible without your knowledge and encouragement.

I am also grateful to Panayotis Antoniadis. You advised me in the beginning of my PhD and continued to support me all these years. Thank you for the amazing time at ETH and for playing an important role in my quest to become a researcher. It was very exciting to work with you.

As a member of NPA and LINCS I was fortunate to be surrounded by remarkable and bright people. Thank you for welcoming me to these labs and for creating a very enjoyable working environment. You are more than colleagues, you are true friends. I will never forget my office friends Mehdi, Nadjet, Yesid, Ahlem, Adisorn, and Tiphail for the great moments and for being around when needed.

I want to thank Andrea Passarella and Walid Dabbous for devoting their time to review my manuscript and for providing valuable remarks on how to improve the manuscript. I would also like to thank Martin May, Anne-Marie Kermarrec, and Sébastien Tixeuil for accepting to be part of my PhD committee. I am truly honored that you have accepted to evaluate my PhD work.

Finally, this thesis is dedicated to my parents, Petru and Nicoleta, and to my brother, Dan-Paul, for their eternal care, love, and support. Va multumesc pentru tot si va iubesc!

Résumé

Il est fondamental de comprendre le comportement des utilisateurs, afin de créer des systèmes de communication efficaces. Dévoiler les interactions complexes entre les utilisateurs dans le monde réel ou en ligne, déchiffrer leurs activités sur Internet, ou bien comprendre la mobilité humaine — toutes les formes des activités – peuvent avoir un impact direct sur la performance d'un réseau de communication. Mais l'observation du comportement de l'utilisateur n'est pas suffisante. Pour transformer l'information en connaissance utile, il faut cependant aller au-delà de l'observation et de l'explication du passé, ainsi que créer des modèles permettant de prédire le comportement.

Dans cette thèse, nous nous concentrons sur le cas des utilisateurs qui consomment du contenu dans leurs trajets quotidiens, en particulier lorsque la connectivité est faible ou intermittente. Nous considérons que les utilisateurs peuvent communiquer entre eux en utilisant l'infrastructure, mais aussi directement, en utilisant les communications opportunistes. Nous proposons de nouvelles perspectives sur la façon d'utiliser de l'information sur le comportement des utilisateurs dans la conception de solutions plus efficaces pour les communications mobiles opportunistes. En particulier, nous mettons en avant le fait que le comportement des utilisateurs, à la fois en termes de consommation de contenu et de contact entre les utilisateurs mobiles, peut être utilisé dans le but d'élaborer des stratégies dynamiques de réplication de données.

On commence par une étude sur les caractéristiques de la consommation de contenu. Notre contribution dans ce domaine est double : tout d'abord, on analyse les caractéristiques sur la publication des nouvelles publiées sur 20minutes.fr, une plate-forme de nouvelles populaire en France ; ensuite, on passe en revue les différents algorithmes de prédiction proposés dans la littérature, on compare la capacité de deux méthodes pour prédire la popularité des articles de presse en ligne. Nous observons qu'un modèle linéaire sur une échelle logarithmique est une solution efficace pour prédire la popularité des nouvelles en ligne. De plus, dans le contexte de classification automatique, nous observons que cette méthode est également une solution efficace pour classer correctement les articles en fonction de leur popularité prédite.

Nous étudions ensuite l'impact d'un modèle capable de prédire la popularité du contenu dans un scénario de communication mobile opportuniste. Nous considérons le contexte de déchargement des données mobiles, où le but est de disséminer du contenu d'une de façon

proactive pendant les périodes d'inactivité, afin de réduire le trafic de données pendant les périodes de pointe. Nous montrons que la capacité de réellement prédire la future demande de l'utilisateur peut améliorer l'effet de dissémination proactive par rapport aux méthodes traditionnelles, qui ne tiennent pas compte que d'une évolution stable de la popularité du contenu. Dans un scénario mobile, les utilisateurs qui partagent un intérêt commun dans le contenu et qui se trouvent dans une proximité physique, peuvent établir des connexions de dispositif à dispositif et ils peuvent, de même, récupérer le contenu directement à partir de leurs voisins.

Nous continuons avec une étude sur la prédiction des contacts entre les utilisateurs mobiles. La mobilité des utilisateurs, représentée comme un système dynamique, n'est pas complètement aléatoire et des modèles peuvent être créés à partir des études sur les mouvements des utilisateurs pendant une certaine période de temps. Mais les contacts entre les utilisateurs mobiles sont une ressource rare : par conséquence, certains utilisateurs vont souvent se rapprocher l'un de l'autre, mais ils ne seront jamais en contact direct. Nous étendons donc la tâche de prédiction pour le cas de k-contact, qui suppose de prédire si les utilisateurs mobiles vont se trouver à une distance d'au plus κ nœuds un de l'autre. En analysant trois traces de contact de la vie réelle, nous observons que, dans un scénario caractérisé par des déconnexions fréquentes, on peut obtenir de meilleures performances lors de la prédiction des nœuds se retrouvant à une plus grande distance les uns des autres, par rapport au cas de contact direct. Pour évaluer l'impact de ces résultats dans un scénario de la vie réelle, nous proposons une expérience de simulation dans laquelle, en combinant les communications mobiles opportunistes avec la prédiction κ-contact, on peut réduire la quantité de trafic utilisé dans la communication de nœuds mobiles avec l'infrastructure cellulaire.

Mots-clés

Réseaux mobiles opportunistes, popularité du contenu web, comportement des utilisateurs, mobilité, réseau de téléphonie mobile, prédiction

Abstract

Understanding user behavior is fundamental in the design of efficient communication systems. Unveiling the complex online and real-life interactions among users, deciphering online activity, or understanding user mobility patterns – all forms of user activity – have a direct impact on the performance of the network. But observing user behavior is not sufficient. To transform information in valuable knowledge, one needs however to make a step forward and go beyond observing and explaining the past to building models that will predict future behavior. In this thesis, we focus on the case of users consuming content on the move, especially when connectivity is poor or intermittent. We consider both traditional infrastructure-based communications and opportunistic device-to-device transfers between neighboring users. We offer new perspectives of how to use additional information about user behavior in the design of more efficient solutions for mobile opportunistic communications. In particular, we put forward the case that the collective user behavior, both in terms of content consumption and contacts between mobile users, can be used to build dynamic data replication strategies.

We first investigate content consumption patterns. Our contribution in this area is two-fold. First, we analyze a large news data set published on `20minutes.fr`, a popular daily newspaper in France. We survey the different prediction algorithms proposed in the literature and compare the ability of two of these methods to predict the popularity of online news articles. We observe that a linear model on a logarithmic scale is an effective solution to predict the popularity of online news. Furthermore, in the context of automatic online news ranking we observe that this method is also an effective solution to correctly rank items based on their future popularity with a performance that can evenly match more customized learning-to-rank algorithms. We study then the practical impact of using a model that can predict content popularity in a mobile opportunistic scenario. We place this in the context of mobile data offloading where the goal is to proactively seed content during idle periods to reduce data traffic during the peak periods. We show that the ability to actually predict future user demand can improve the benefit of proactive seeding for a mobile opportunistic data offloading solution compared to traditional methods that consider a stable evolution of content popularity.

In a mobile scenario, users who share common interest in a content and are within physical proximity, can establish the device-to-device connections and retrieve content directly

from their neighbors. We study the predictability of human contacts. User mobility, represented as a highly dynamic system, is not completely random and patterns can be learned after studying user movement for a certain period of time. But contacts between mobile users are a scarce resource, as some users will often come close to each other but never in direct contact. We thus extend the prediction task to the multi-hop contact case – predict if mobile users will find themselves at a distance of at most κ-hops from one another. By analyzing three real-life contact traces we observe that, in a mobile scenario characterized by frequent disconnections, one can obtain better performance when predicting that nodes will find themselves at a greater distance from one another compared to the direct contact case. To assess the impact of these findings in a real-life scenario, we propose a simulation experiment in which, by combining mobile opportunistic communications with κ-contact prediction, one can reduce the amount of traffic used in the communication of mobile nodes with the cellular infrastructure.

Key Words

Mobile opportunistic networks, web content popularity, user behavior, mobility, cellular network, prediction

Table of contents

1 Introduction **15**
- 1.1 Context and motivation . . . 15
- 1.2 Global scenario and research challenges . . . 17
- 1.3 Contributions of this thesis . . . 19
  - 1.3.1 A survey on predicting the popularity of web content . . . 19
  - 1.3.2 Predicting the popularity of online news . . . 20
  - 1.3.3 Proactive seeding based on content popularity prediction . . . 20
  - 1.3.4 Predicting κ-contact opportunities between mobile users . . . 20
- 1.4 Thesis outline . . . 21

2 A survey on predicting the popularity of web content **23**
- 2.1 Introduction . . . 23
- 2.2 Domains . . . 24
- 2.3 Performance measures . . . 27
  - 2.3.1 Numeric prediction . . . 27
  - 2.3.2 Classification . . . 29
- 2.4 A classification of web content popularity prediction methods . . . 29
  - 2.4.1 Single domain . . . 29
    - 2.4.1.1 Before publication . . . 30
    - 2.4.1.2 After publication . . . 30
  - 2.4.2 Cross domain . . . 31
- 2.5 A survey on popularity prediction methods . . . 31
  - 2.5.1 Single domain . . . 31
    - 2.5.1.1 Before publication . . . 31
    - 2.5.1.2 After publication - Aggregate behavior . . . 32
    - 2.5.1.3 After publication - Individual behavior . . . 38
  - 2.5.2 Cross domain . . . 39
- 2.6 Selecting the right features . . . 44
- 2.7 Factors that influence content popularity . . . 46

2.8 Predictive proactive seeding: an application of web content popularity prediction . . . 47
2.9 Conclusions . . . 47

3 Predicting the popularity of online news articles 49
3.1 Introduction . . . 49
3.2 Background . . . 50
3.3 Global statistics . . . 51
3.3.1 Online news data collections . . . 51
3.3.2 News articles lifetime . . . 52
3.3.3 Distribution of popularity . . . 54
3.4 Predicting the popularity of online news articles . . . 56
3.4.1 Popularity predictions methods . . . 56
3.4.2 Popularity prediction accuracy . . . 58
3.5 Ranking news articles based on popularity prediction . . . 60
3.5.1 Methodology . . . 60
3.5.2 Ranking methods . . . 62
3.5.3 Ranking performance . . . 63
3.5.4 An alternative to learning to rank algorithms . . . 64
3.6 Conclusions . . . 66

4 Predictive proactive seeding for mobile opportunistic data offloading 67
4.1 Introduction . . . 67
4.2 Background . . . 68
4.3 Global scenario . . . 69
4.4 Proactive seeding in mobile opportunistic networks . . . 71
4.4.1 Premise for effective proactive seeding . . . 71
4.4.2 Proactive seeding strategies . . . 72
4.5 Evaluation . . . 74
4.5.1 Simulating user behavior . . . 74
4.5.2 Simulation scenario . . . 76
4.5.3 Results . . . 77
4.6 Conclusion . . . 78

5 Beyond contact predictions in mobile opportunistic networks 81
5.1 Introduction . . . 81
5.2 Background . . . 82
5.3 Vicinity and data sets . . . 83
5.3.1 Beyond contact relationships . . . 83
5.3.2 κ-vicinity, κ-contact, and κ-intercontact . . . 84
5.3.3 Data sets . . . 85
5.4 Pairwise relationships under the κ-contact case . . . 86
5.4.1 Pairwise minimum distance . . . 86
5.4.2 Analyzing the distribution of pairwise distance . . . 87

5.4.3 The stability of κ-contact relationships . . . 89
5.5 Predicting κ-contact encounters . . . 90
5.5.1 Dynamic graph representation . . . 90
5.5.2 κ-contact prediction problem . . . 90
5.5.3 The effect of time-window duration and past data . . . 92
5.5.4 κ-contact prediction results . . . 94
5.6 Practical implications . . . 95
5.7 Conclusions . . . 97

6 Conclusions and future work **99**
6.1 Summary . . . 99
6.2 Looking ahead . . . 100
6.2.1 Improving the quality of the prediction . . . 100
6.2.2 Smart proactive seeding . . . 102
6.2.3 Predicting spatiotemporal contacts . . . 102
6.2.4 Mobile opportunistic data offloading engine . . . 103

A Résumé en francais **105**
A.1 Contexte et motivation . . . 105
A.2 La problématique . . . 107
A.3 Contributions de cette thèse . . . 110
A.3.1 Une synthèse sur les algorithmes de prédiction de la popularité du contenu web. . . . 110
A.3.2 Prédire la popularité des articles . . . 112
A.3.3 Pré-téléchargement du contenu fondé sur la prédiction de la popularité du contenu . . . 112
A.3.4 Prédire les événements κ-contact entre les utilisateurs mobiles . . . 113
A.4 Conclusions . . . 115
A.5 Perspectives . . . 116
A.5.1 Améliorer la qualité de la prédiction de popularité . . . 116
A.5.2 Pré-téléchargement intelligent . . . 118
A.5.3 Prédire les contacts spatio-temporelles . . . 119
A.5.4 Moteur pour le délestage opportuniste de trafic mobile de données . 119

B List of Publications **121**

List of figures **121**

List of tables **126**

References **127**

Chapter 1

Introduction

1.1 Context and motivation

In the past years we have assisted at two emerging trends in the evolution of Internet. First, the production of online content has evolved from the hands of a few (e.g., traditional media organizations) to the general public. Stimulated by the technological progress brought by Web 2.0 platforms and the explosive growth of social networking sites, online users are now able to create and share content on their own – and often to rival professional content creators. Then, we assist at an important shift towards mobile web access. Equipped with better web-enabled mobile devices (smaller, cheaper, and with improved functionalities), and under a better broadband mobile coverage, data-hungry consumers crave to consume content at any moment of time. One good example of the rapid proliferation of mobile devices is reflected by two photos, taken eight years apart, and representing people gathering at St. Peter's Square (Figure A.1).

These two trends have stimulated users' needs to connect anywhere and anytime – to other users and to information – and to create, consume, and share content at an unprecedented pace. For example, every minute, users around the world send more than 300,000 tweets [1], share more than 680,000 pieces of content on Facebook [2], and upload 100 hours of video on YouTube [3]. And this increase is not an isolated trend. As it turns out, the global volume of data has been increasing at a rate of 50% per year and there has been a 40-fold increase compared to 2001 [4]. And, while storing this huge amount of information still seems manageable (e.g., it costs 600$ to store all the music in the world [4]), providing the infrastructure to make content always available is a challenging objective.

Telecom operators strive under the increasing mobile data consumption. Traditionally, when the network capacity reached critical limits, operators relied on certain technical



(a) 2005 (b) 2013

Figure 1.1: People gathering in St. Peter's Square eight years apart (Source: NBC[1]).

solutions: acquire additional spectrum, deploy more cell sites, or switch to the latest mobile communication standard (e.g., LTE Advanced). But nowadays, there is a general believe that these solutions may not hold to the challenges of the years to come. Spectrum is a finite resource (and also the more and more expensive), spectrum efficiency is quickly reaching its limits, and installing additional cell sites has a significant cost. As a result, new solutions have been proposed to cope with the expected traffic increase. Mobile data offloading is one attractive solution that enables telecom operators to shift part of the traffic from the cellular networks to alternative low-cost networks [5]. In this context, Wi-Fi hotspots are a valuable resource as they are widely available, have a low cost, and good data rates [6]. Femtocells are also a promising alternative that allow a better utilization of the available spectrum [7]. However, the offloading potential of these solutions may not sustain the rate of data traffic increase and novel solutions are needed.

Opportunistic networks have been recently been proposed as an appealing solution to offload content with non-real time constrains. Instead of using the cellular network infrastructure, mobile users can retrieve content from collocated peers that share common interest (and are willing to participate in the network operation). This opens new directions on how people can generate and consume content on-the-go, but the design of efficient communication protocols in opportunistic networks is challenging (due to user mobility, it is difficult to make assumptions about the existence of a path between two nodes) and it depends in great part on the capacity to understand user behavior. *Thus, capturing the dynamics of user behavior, discovering regularities, and learning predictive patterns becomes vital in the design of opportunistic network communication protocols.*

So far, most studies on user behavior have focused on a better understanding of user mobility and connectivity patterns. This includes insights about the duration of contacts (and inter-contacts) between mobile users [8], the periodicity of human encounters [9], or in understanding the underlying social structures (physical and online) that may explain

[1] http://instagram.com/p/W2FCksR9-e/

human mobility patterns [10, 11]. Studying user mobility is essential, but in the complex environment under which mobile users operate there are other aspects about user behavior, equally important, that can be exploited.

The objective of this thesis is to offer new perspectives on how to use additional aspects about mobile users' behavior in the design of efficient mobile opportunistic data offloading solutions. In particular, we address from a different angle the problem of predicting contacts between mobile users and we put forward the case that the design of efficient mobile opportunistic data offloading strategies should pay attention not only to the mobility aspect but also to what users consume.

1.2 Global scenario and research challenges

The global scenario considered throughout this work and illustrated in Figure A.2 is composed of three main entities: a content producer, a telecom operator, and a group of collocated mobile users. The content producer is located on the Internet and periodically publishes content for a group of collocated mobile users. The telecom operator provides the infrastructure for the communication between mobile users and the content producer. Finally, we consider a group of collocated mobile users that communicate with the content producer using the cellular infrastructure and can also communicate directly between each other using device-to-device communication techniques (e.g., Bluetooth, Wi-Fi direct).

This network environment corresponds to a certain urban area (e.g., university campus or commercial center) populated by mobile users that show localized, geographical and temporal, data access patterns. In this context, we distinguish two possible strategies for the mobile users to access content on-the-go. In the classical approach, users rely on the services provided by the telecom operator and individually retrieve content using the infrastructure. But, given that the temporal-geographical correlation of user requests, this approach seems outdated and inefficient as there is a high chance that content could directly be retrieved from collocated mobile users. The alternative solution is to rely on opportunistic communications (if the geographic area is densely populated to provide good means of communication between nodes) and give mobile users the possibility to communicate directly and share the cache space of collocated mobile users.

Different solutions have been proposed in the latest years, which consider the coexistence of infrastructure with the opportunistic networks communications to decide when and where (to which users) to proactively seed content in order to reduce the communication of mobile users with the infrastructure [12, 13]. But the current implementations of opportunistic networks are rather myopic in the sense that they do not fully benefit from the potential knowledge about user behavior. For example, by tracking the content request patterns,

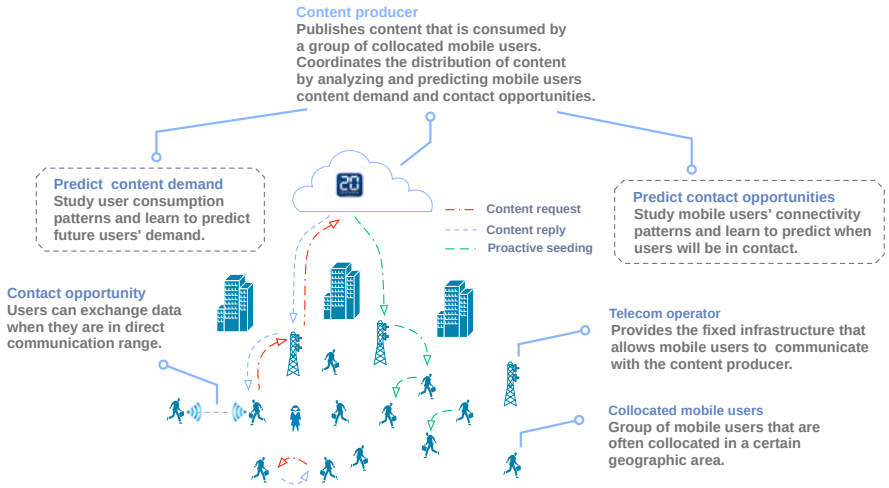


Figure 1.2: The global scenario considered throughout this work and composed of a content producer, located on the Internet, a telecom operator that provides the infrastructure for the communication between mobile users and the content provider, and a set of collocated mobile users.

one can predict content popularity and decide to proactively seed part of the content to better satisfy future user demand. Furthermore, even greater benefits could be attained by tracking the connectivity patterns between mobile users and predicting the evolution of the network topology.

The problem addressed in this thesis is how to design more effective mobile opportunistic data offloading solutions based on a global knowledge about users content requests and connectivity patterns. In particular we address the following two problems:

- **Problem 1**

What content to seed? Given the great amount of content published on a daily basis, the skewed distribution of users' interest, and the non-stationary content popularity (the popularity of a piece of content evolves over time) decide on what content to seed and the number of replicas to cope with the future users' demand. This allows one to build adaptive proactive seeding techniques that are consistent with the dynamic evolution of content popularity.

Approach. We undertake this problem using the following steps. First, we want to understand to what extend the popularity of web content can be predicted. We take online news articles as a use-case example and study the popularity of articles published on two popular news platforms from France and Netherlands. We look into the various methods that have

been proposed to predict the popularity of web content, select the ones that are adapted to our case, and study their effectiveness to predict the popularity of news articles. We then assess the impact of these findings as a solution to improve the benefit of proactive seeding solutions in the context of mobile data offloading.

• **Problem 2**

Where to seed? Given the opportunistic nature of human encounters that are, to a certain extent, predictable decide on how to better organize the seeding process by choosing where (to which nodes) to seed content.

Approach. We analyze various human-based contact traces and study the predictability of human contacts. Given the rather unpredictable nature of these relationships we extend our analysis to the κ-contact case – predict if users will find themselves at a distance of at most κ-hops from one another. To assess the impact of these findings in a real-life application, we propose a simulation experiment in which, by combining mobile opportunistic communications with κ-contact prediction one can reduce the amount of traffic used in the communication of mobile nodes with the infrastructure.

1.3 Contributions of this thesis

1.3.1 A survey on predicting the popularity of web content

When studying the popularity of web content there is no clear evidence that there is a prediction model that could be applied to every possible scenario nor that the creation of a generic prediction model is a feasible objective. The reasons are that the prediction outcomes are influenced by the type of online content, the site's framework, and the availability of predictive information. Thus, in the field of social media various popularity prediction methods have been proposed and evaluated on different types of web content.

The first contribution of this thesis is a survey on the current state-of-the-art in web content popularity prediction methods. This domain has become an active area of research and, while still in an incipient phase, a large number of prediction methods for different types of web content have been proposed in the latest years. To structure the existing prediction methods we propose a classification based on the type of information used in the prediction process. We report the performance of the different prediction methods, present the features that have showed good predictive capabilities, and reveal factors known to influence content popularity.

1.3.2 Predicting the popularity of online news

We study two popular online news platforms from France and Netherlands to understand how articles are consumed by online readers. By exploring these two large data sets (one that covers a 4-year period and another one an 8-month period) we observe that news articles have a very short lifespan and that the volume of comments per article can be described by a power-law distribution. Using these data sets we analyze the predictive power of two popularity prediction methods. Our results indicate that a linear model on a logarithmic scale provides the most accurate results in predicting the popularity of online news articles.

In the context of automatic online news ranking we evaluate the ranking effectiveness of the popularity prediction methods and show that a linear model on a logarithmic scale is an effective method for online news ranking. We compare the performance of these methods with learning to rank algorithms and show that for this ranking problem, popularity prediction methods could successfully replace more customized learning to rank algorithms.

1.3.3 Proactive seeding based on content popularity prediction

We propose and evaluate the benefit of using a proactive seeding strategy, based on web content popularity prediction method, as a solution for mobile opportunistic data offloading. Compared to traditional strategies that consider a rather stable evolution of content popularity over time, the strategy used in this case is to actually predict future content demand and adjust the proactive seeding decisions accordingly.

To evaluate the benefit of this solution in a real-life deployment, we proposed a simulation scenario that reproduces, to a certain extent, the mobility and data request characteristics of a group of mobile users. In this scenario, the objective is to reduce the amount of traffic that the mobile users create during the day by preloading content when the network is less loaded. We show through simulation that proactive seeding can have a greater impact for mobile data offloading if the decision of what content to replicate is based on an algorithm that predicts future users' requests.

1.3.4 Predicting κ-contact opportunities between mobile users

We study the problem of predicting future connectivities between mobile users. Given the rather limited ability to predict contacts between mobile users we extend our prediction scope to the κ-contact case – predict if users will find themselves at a distance of at most κ-hops from one another. Using a supervised prediction framework we analyze the predictability of κ-contacts on three real-life contact traces, and observe that one can attain better performances when predicting that users will not be in direct contact but in the

nearby vicinity. To assess the impact of these findings in a real-life deplyment, we propose a simulation experiment in which, by combining mobile opportunistic communications with κ-contact prediction one can reduce the amount of traffic used in the communication of mobile nodes with the infrastructure. Our results suggest that services benefiting from contact predictions can efficiently exploit the predictable nature of κ-contacts.

1.4 Thesis outline

The remainder of this thesis is organized as follows. We begin with a classification and a presentation of the methods used to predict the popularity of web content (Chapter 2). Then, in Chapter 3, we analyze the ability to predict the popularity of online news articles. In Chapter 4 we propose the application of popularity prediction methods in proactive seeding and evaluate the impact of this approach as a solution for mobile opportunistic data offloading. We study user mobility characteristics and analyze the predictability of human encounters in Chapter 5. We conclude in Chapter 6 with a discussion of some possible future directions.

Chapter 2

A survey on predicting the popularity of web content

2.1 Introduction

In the digital world, web content has become the main attraction. Whether it is useful information and entertainment to Internet users or business opportunity for marketing companies and content providers, web content is an asset on the Internet. At the same time, the growth in social media innovation, the ease of content creation and low publishing costs, has created a world saturated with information. For example, every minute, users around the world send more than 300,000 tweets [1], share more than 680,000 pieces of content on Facebook [2], and upload 100 hours of video on YouTube [3]. Yet, the online ecosystem adheres to a "winner-take-all" society: the attention is concentrated on a few items while the majority remains unknown. In this context, finding the web item that will be popular becomes of utmost importance. Online users, flooded by information, can reduce the clutter and focus their attention – the most valuable resource in the online world – on the most relevant information. In a world where companies spend up to 30% of their money in online marketing [14], spotting early on the next rising star of the Internet can maximize their revenues through better add placement. Moreover, given the ever-growing consumer Internet traffic, content-distribution networks can rely on popularity prediction methods to proactively allocate resources according to future user demands.

The term *web content* is effectively generic and it broadly defines any type of information on a web site. It can refer both to the subject of the information and the individual object used to deliver the information. In this dissertations we define web content as any individual item, publicly available on a web site and which contains a measure that reflects a certain

interest shown by an online community.

The notion of popularity for a web content is subtle, beyond the usual number of page views. Before Web 2.0, the competition was on counting "eyeballs", but now, with the growing prevalence of social media platforms, there are new indicators that reflect user interest. In response to content publication, users can now provide a direct feedback, thought comments and ratings, or further share it in their online social circles (through, for example, Facebook, Twitter, or Digg). These alternative metrics capture user engagement at a much deeper level and provide valuable information complementary to view counts: rating improves the quality of publications, comments increase the time spent on a web page (which impacts the advertising revenues), and sharing gives content a larger notoriety. In addition, these metrics capture distinct user habits as users have different preferences (to comment, rate, or share) in their post-click actions [15–18]. In this context, studying these metrics individually or how they relate to each other [19, 20] provides a wider and better perspective of what popularity actually means.

Predicting the popularity of web content is a challenging task. First, different factors known to influence content popularity, such as the content quality or its relevance to users, are difficult to measure. Then, other factors, such as the relationship between events in the physical world and the content itself are hard to capture and used in a prediction model. Moreover, at a microscopic level, the evolution of content popularity may be described by complex online interactions and information cascades, which are difficult to predict [21–23].

Predicting the popularity of web content has become an active area of research and, while still in an incipient phase, a large number of prediction methods for different types of web items have been proposed in the latest years. As a first step to actually predicting the popularity of news articles, to have a better understanding on the challenges and the existing solutions, in this chapter we look from a general point of view to the problem of predicting the popularity of web content. We review the current state of research in this field, propose a classification that allows us to structure the different prediction methods, and briefly describe the main prediction methods.

2.2 Domains

The consumption of web content is spread across multiple domains and a variety of items. Some of the most popular types of web items studied so far include: user-generated videos, which account for a great percent of Internet traffic [24]; news articles, massively diffused through social networking sites [25] and heavily consumed on mobile devices [26]; stories published on social news aggregators, which provide an even greater exposure to the most popular items on the Internet; and items (comments, photos, or videos) published on

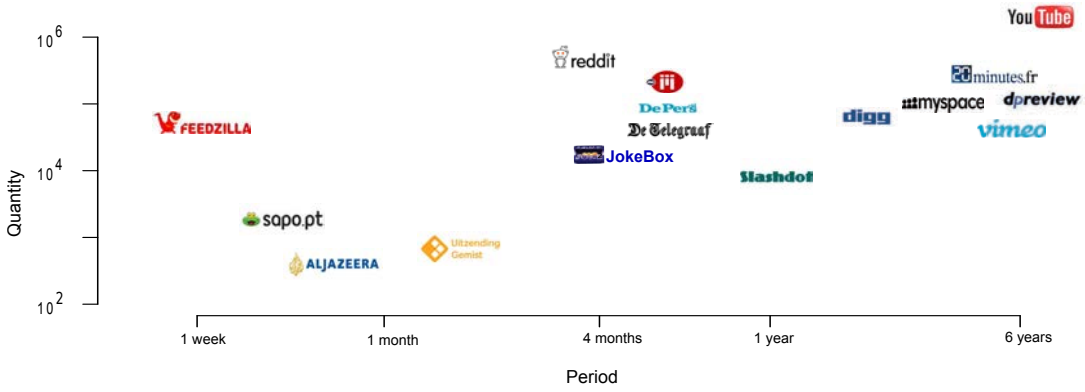


Figure 2.1: Data sets used as case-studies to evaluate the performance of prediction methods. On a log-log scale we depict the total number of items and the cumulative time period covered by each data set (using a **w**eekly, **m**onthly, **y**early demarcation).

social networking sites, the most popular platforms to share information and allow user participation on a global scale. Examples of the variety of web items, gathered from different platforms and used in the context of popularity prediction, are illustrated in Figure 1, together with information about the number of items and the time period covered by each data set.

Online videos. YouTube, the world's largest video sharing platform with 100 hours of upload per minute [3] and more than 1 trillion worldwide views per year [27], has been the main focus of existing studies. The site's content, with more than 200 million unique videos, covers a broad range of topics and is sustained by a large and active online community [28]. Studying the popularity of YouTube videos is challenging given the ever-growing number of videos, the many features that the platform provides (e.g., video recommendations, internal search, online social networking), and the limitations associated with the retrieval of a representative sample of videos [29].

The popularity of YouTube videos (commonly expressed by the number of views in research studies) follows a heavy-tailed distribution that, depending on the data set and the method used to fit the distribution, can be described by Zipf [30,31], power-law with exponential cut-off [32], Weibull [28], or Gamma distributions [33]. But video access frequency over time is highly non-stationary. From a high-level point of view, the popularity growth of videos over time can be represented through power-law or exponential distributions [34]. A more fine-grained analysis exhibits even more complex and diverse patterns. For instance, Crane and Sornette found that, while the activity around most YouTube videos can be described by a Poisson process, many videos reveal similar activity around the peak

period that can be accurately described by three popularity evolution patterns [35]. Similar temporal evolution patterns have been observed by Figueiredo [36] and even more complex shapes have been discovered by Gorsun et al. [29].

In addition to YouTube, the popularity of videos published on other platforms has been studied (e.g., Daum [32], Dutch TV [34], DailyMotion [37], Vimeo [38]), but on a smaller scale, and no significant differences have been signaled.

Online news. The primary source of information in the digital world, news are created in large numbers and massively diffused through online social networks [25]. Compared to videos, which catch users' attention for a longer period of time, the interest in news articles fades quickly, within days after publication [19,39]. The popularity of online news, frequently expressed by the volume of comments (the number of views are rarely disclosed by news platforms), also follows a heavy tail distribution, described by power-law [40, 41] or log-normal [42] distributions.

Social bookmarking sites. The third major type of content analyzed so far is stories posted on social bookmarking sites such as Digg [43–45], Slashdot [46], or Reddit [45, 47]. Content published on these sites experiences an even greater rate of change with stories reaching their attention peak in the first six hours after publication and being completely saturated within one day [43]. Prediction becomes even more difficult in this setting given the complex interactions between users [48, 49] and the promotion algorithm based on collective user opinion [50, 51]. Similar to other types of online items, stories published on these sites are described by a heavy-tailed distribution of popularity that is best represented by a Weibull [45] or log-normal distributions [43, 44, 52].

Social networking services. Designed with the idea of facilitating interactions among people on the Internet, these sites allow users to build and maintain online social relationships with people that share common interest, background, or real-life relationships. While there are different types of social networking services, the most popular ones today, are the ones focused on content sharing. Microblogs, such as Twitter and Weibo, are a specific type of social networking services and have been extensively studied. These platforms are probably the most dynamic representation of social media. Users create and share information in the form of short messages (known as tweets) containing up to 140 characters. When users post a (re)tweet it becomes visible to all its followers (members of the social group). Content can easily spread though the social graph as followers can further share the content (known are retweeting) to their own list of followers. Two metrics have been used to measure the popularity of a tweet: the number of impressions (number of online users who viewed a certain tweet), or most commonly, the number of retweets. As other type of online items the popularity of a tweet (in form of retweet counts) also follows a

power-law distribution [53–55].

Tweets are probably one the most ephemeral type of online content as they become popular very fast and they quickly die out. For example, studies conducted on Tencent Weibo found out that an insignificant number of tweets get retweeted after one day [55]. Similar, a study on Twitter revealed that most tweets receive half of their retweets within one hour after publication [56]. Useful predictions thus need to be done in the order of minutes after a tweet post. In addition, as content is shared between users, large cascades may form, which have a very unpredictable nature.

In addition to these main categories, other types of web items have been considered for popularity prediction tasks such as threads on discussion forums (DPReview, MySpace [57]), movie ratings on IMDb [58], an interactive video sharing application (Zync) [59], and a joke sharing application (JokeBox) [60].

2.3 Performance measures

To provide a more explicit description of the prediction algorithms, let us introduce the terminology and the measures used to evaluate the efficiency of a prediction method.

Terminology. Let $c \in C$ be an individual item from a set C observed during a period T. We use $t \in T$ to describe the age of a web item (duration since it was published) and mark two important moments: indication time t_i, representing the time we perform the prediction; reference time t_r, the moment of time when we want to predict the popularity. Let $N_c(t_i)$ be the popularity of c from the time it was published until t_i and let $N_c(t_r)$ be the value that we want to predict, i.e., the popularity at a later time t_r. We define $\widehat{N}_c(t_i, t_r)$ the prediction outcome: the predicted popularity of c at t_r using the information available until t_i. Thus, the better the prediction, the closer $\widehat{N}_c(t_i, t_r)$ is to $N_c(t_r)$.

Evaluation. We distinguish two prediction goals: (i) Numeric prediction – predict the exact popularity that an individual item will generate, (ii) Classification – predict the popularity range that a web item is most likely to fall in.

2.3.1 Numeric prediction

There are different ways to assess the efficiency of a numeric prediction [61]. Mean squared error (MSE – Equation 2.1) is used to report the average of the squared errors. By taking the square root of MSE, one can express the error in the same dimension as the estimated value (RMSE – Equation 2.2). One important limitation of squared errors is that they put too much weight on the effect of outliers, and in this case reporting the absolute errors is a good alternative (MAE – Equation 2.3).

Absolute errors can be meaningfully interpreted if one knows the range of the actual popularity values. Otherwise, a good way of expressing the error is though relative errors such as the Mean Relative Error (MRE – Equation 2.4) and Mean Relative Squared Error (MRSE – Equation 2.5). Relative measures are also useful to compare the efficiency of prediction algorithm across studies, as in most cases the popularity values have widely different ranges (e.g., the number of views on YouTube is several orders of magnitude greater then the number of comments on a news web site). A special attention should be paid when using these error measures for zero-inflated variables as the relative error is undefined when the actual value is zero.

Another way of expressing the error is through the Relative Squared Error (RSE – Equation 2.6), Root Relative Squared Error (RRSE – Equation 2.7), and Relative Absolute Error (RAE – Equation 2.8). The error in this case is expressed relative to the performance of a simple predictor, the average of the actual values (computed on the training data set).

Finally, a rather different way of reporting the quality of a numeric prediction is through the correlation coefficient or the coefficient of determination (R^2). Compared to the previous measures, which show how the estimated values diverge from the actual ones, these evaluation criteria can only express the degree of linear association between the two variables (predicted and actual values).

$$\text{MSE} = \frac{1}{|C|}\sum_{c}\left(\widehat{N}_c(t_i,t_r) - N_c(t_r)\right)^2. \tag{2.1}$$

$$\text{RMSE} = \sqrt{\frac{1}{|C|}\sum_{c}\left(\widehat{N}_c(t_i,t_r) - N_c(t_r)\right)^2}. \tag{2.2}$$

$$\text{MAE} = \frac{1}{|C|}\sum_{c}\left|\widehat{N}_c(t_i,t_r) - N_c(t_r)\right|. \tag{2.3}$$

$$\text{MRE} = \frac{1}{|C|}\sum_{c}\left|\frac{\widehat{N}_c(t_i,t_r) - N_c(t_r)}{N_c(t_r)}\right|. \tag{2.4}$$

$$\text{MRSE} = \frac{1}{|C|}\sum_{c}\left(\frac{\widehat{N}_c(t_i,t_r) - N_c(t_r)}{N_c(t_r)}\right)^2. \tag{2.5}$$

$$\text{RSE} = \frac{\sum_c\left(\widehat{N}_c(t_i,t_r) - N_c(t_r)\right)^2}{\sum_c\left(\widehat{N}_c(t_r - \overline{N}(t_r)\right)^2}. \tag{2.6}$$

$$\text{RRSE} = \sqrt{\frac{\sum_c\left(\widehat{N}_c(t_i,t_r) - N_c(t_r)\right)^2}{\sum_c\left(\widehat{N}_c(t_r - \overline{N}(t_r)\right)^2}}. \tag{2.7}$$

$$\text{RAE} = \frac{\sum_c \left| \widehat{N}_c(t_i, t_r) - N_c(t_r) \right|}{\sum_c \left| (\widehat{N}_c(t_r - \overline{N}(t_r) \right|}. \tag{2.8}$$

2.3.2 Classification

The prediction task can also be treated as a classification problem, where, assuming that the popularity range is known, one can split this interval in k classes (non-overlapping popularity ranges). Thus, given the k possible outcomes the prediction goal is to correctly predict the popularity class.

Various metrics are available to evaluate the quality of a classification [61,62]. *Accuracy*, one of the most reported metric, is used to express the proportion of correctly classified instances. This measure is nevertheless inappropriate when dealing with highly imbalanced classes, which can often be the case with the popularity of web items (characterized by a heavy-tail nature). For example, a possible experiment could be to learn a classifier that predicts which videos will get more that 10^6 views on YouTube - a "small" class (1%) according to a recent study [63]. A simple rule, that decides that all videos receive less than 10^6, will correctly predict 99% of the cases. Thus, a good level of accuracy is obtained without even learning any prediction rule on how to detect the popular objects. To measure the performance of the classifier on a "small" class, a good alternative is to use *precision*, *recall*, or *F-score* (harmonic mean between *precision* and *recall*). But *F-score* measures the performance of a classifier for only one class. To report the aggregate performance over multiple classes, a good solution is to use *macro-average* measure (average F-score over all k classes).

2.4 A classification of web content popularity prediction methods

To structure the different popularity prediction methods, we propose a classification that groups the methods according to the type and granularity of the information used in the prediction process (Figure 2.2). We further organize the subsequent chapters based on this classification.

2.4.1 Single domain

We define a domain as the web site where an individual item resides, regardless if it has been created or shared from an external source (e.g., news shared on social bookmarking

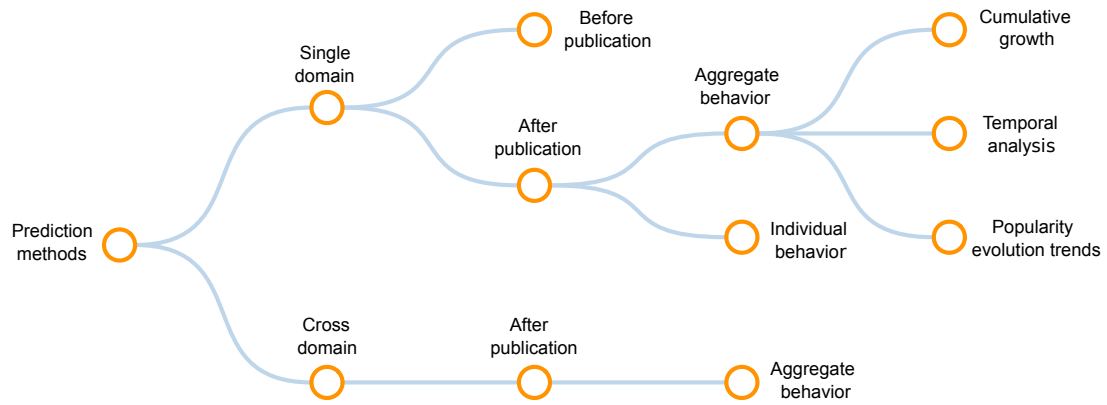


Figure 2.2: A classification of web content popularity prediction methods.

sites). Methods under this category make forecasts about popularity using only the local information about the web item.

2.4.1.1 Before publication

One of the most challenging objectives is to predict the popularity of a web content before its publication, relying only on content metadata or the social structure of the publisher.

2.4.1.2 After publication

The alternative is to include in the prediction model data about the attention that one item receives after its publication.

Aggregate behavior. A common approach is to deduce future content popularity from the aggregate early users' reactions. This can further be separated in three main categories:

- Study the *cumulative growth* of attention, i.e., the amount of attention that a web item receives from the moment it was published until the prediction moment.
- Perform a *temporal analysis* of how content popularity evolved in time until the prediction moment.
- Use clustering methods to find *popularity evolution trends*.

Individual behavior. Instead of treating each user action equally, one may further refine the prediction model by taking into account individual user behavior.

2.4.2 Cross domain

Explaining popularity from the perspective of single domain is limited due to the diverse and complex user interactions between different platforms. Methods under this class draw conclusions by extracting and transferring information across web domains.

2.5 A survey on popularity prediction methods

Several popularity prediction methods have been proposed in the last decade, from simple linear regression functions to complex frameworks that mine and build knowledge from social media. We describe these methods according to the proposed classification and explain how these methods perform on different types of web content. A summary of these methods is then presented in Table 2.1.

2.5.1 Single domain

In the vast majority of cases, prediction methods rely entirely on the information available on the site where the content has been published.

2.5.1.1 Before publication

Predicting the popularity of an item before publication is particularly useful for web items with short lifespan. News articles, which are time-sensitive by nature, fall under this category and have been analyzed in two studies [64, 65].

Tsakias et al. addressed this problem as a two-steps classification problem: predict if news articles will receive comments and if they do, if the number of comments will be high or low [64]. The prediction method used in this case has been a RandomForest classifier trained on a large number of features (textual, semantic, and real-world). Using several online news sources the authors showed that one can accurately predict which articles will receive comments and observed that the performance degrades significantly when trying to predict if the volume of comments will be high or low.

Bandari et al., using the number of tweets as an indicator of news popularity, formulated the prediction task both as a numeric and a classification problem [65]. Predicting the exact popularity of news articles, even under various regression methods (linear, k-nearest-neighbors and support vector machines (SVM) regression) showed modest results, being able to explain only 34% of the variability in the observed popularity ($R^2 = 0.34$). Predicting ranges of popularity has proved to be more effective, with an accuracy of 84% when identifying articles that would receive a small, medium, or large number of tweets.

2.5.1.2 After publication - Aggregate behavior

The methods under this category have been used to predict the popularity of web items based on the aggregate user attention received early after publication.

Cumulative growth. One of the first solutions, which was used to model the popularity of Slashdot stories, was proposed by Kaltenbrunner et al. [46]. The model, which we will refer to as `growth profile` (we adopt the terminology used in [43]), assumes that, depending on the time of the publication, news stories follow a constant growth that can be described by the following function:

$$\widehat{N}_c(t_i, t_r) = \frac{N_c(t_i)}{P(t_i, t_r)}, \tag{2.9}$$

where $P(t_i, t_r)$ is a rescaling parameter and represents the average growth of a story from t_i to t_r

$$P(t_i, t_r) = \frac{1}{|C|} \sum_c \frac{N_c(t_i)}{N_c(t_r)}. \tag{2.10}$$

The method has been tested on a large corpus of Slashdot stories and showed reasonable performance in predicting the popularity of news stories using the aggregate user reactions in the first day after publication (average MRE of 36%).

Describing future popularity as a linear relationship of popularity at earlier stages has also been done by Szabo and Huberman under the `constant scaling` model [43]:

$$\widehat{N}_c(t_i, t_r) = \alpha_2(t_i, t_r) N_c(t_i). \tag{2.11}$$

Parameter α is computed in such a way that the model minimizes MRSE (by setting the first derivate to zero) and is described by the following expression:

$$\alpha(t_i, t_r) = \frac{\sum_c \frac{N_c(t_i)}{N_c(t_r)}}{\sum_c \left[\frac{N_c(t_i)}{N_c(t_r)}\right]^2}. \tag{2.12}$$

Szabo and Huberman also observed a strong correlation between the popularity of an item early after its submission and its popularity at a later stage and proposed a logarithmically transformed linear regression model (`log-linear`) expressed as

$$\widehat{N}_c(t_i, t_r) = \exp\left(\ln N_c(t_i) + \beta_0(t_i, t_r) + \frac{\sigma_0^2(t_i, t_r)}{2}\right). \tag{2.13}$$

For the coefficients of Equation 3.1, β_0 is computed on the training set using maximum likelihood parameter estimation on the regression function $\ln N_c(t_r) = \beta_0(t_i, t_r) + \ln N_c(t_i)$ and σ_0^2 is the estimate of the variance of the residuals on a logarithmic scale.

This method showed good predictive performance on several data sets: Digg stories [43], YouTube videos [43], articles published on a French news platform [66], and Dutch online news articles [42]. For example, Tsagkias et al. observed that, by using the number of comments in the first ten hours after the publication, one can attain good performances in predicting the final volume of comments (average MRSE of 20%) [42].

A different approach has been proposed by Lee et al. [57]. Instead of predicting the exact amount of attention the authors study the possibility to predict if web items will continue to receive attention from online readers after a certain period of time. The prediction model used for this problem (Cox proportional-hazards regression) is a widely used method in `survival analysis` that allows one to model the time until an event occurs (a typical event is "death", from which the term survival analysis is derived). While the main utilization of this method could be to predict the lifetime of a web content, by changing the definition of an event, the method can also be used also for popularity prediction tasks. The solution proposed by Lee et al. is to consider as event the time when a web content will reach a popularity above a certain threshold. The performance of this method has been studied on threads from two online discussion forums (DPreview and MySpace) with popularity expressed as the volume of comments per thread. Using different statistics related to the users' comment arrival rate the authors showed that, by observing user activity in the first day after publication, the method can detect with an 80% accuracy threads that receive more than 100 comments.

Regression-based methods have been frequently used for this prediction task. Tatar et al. used a simple linear regression based on the early number of comments to predict the final number of comments for news articles [67]. The authors observed that there is no significant improvement when using specialized prediction models as a function of the category and the publication hour of an article. Marujo et al. studied the problem of predicting the number of clicks that news stories will receive during one hour. Various prediction methods have been tested (multiple linear regression, regression-based trees, bagging, and additive regression) using different features extracted from a news web platform. Their results indicate that by combining different regression algorithms one can obtain fairly good results ($MRE = 12\%$) in predicting the number of clicks received by news articles during one hour. Cho et al. used a linear model on a logarithmic scale to predict popularity ranges for political blog posts [68]. They show that, by looking at the number of page views in the first 30 minutes, one can classify articles in three classes of popularity with 86% accuracy. A different approach, described more in detail in Chapter 3, was proposed by Tatar et al. who studied the performance of two popularity prediction methods (linear-log and constant scaling) to rank news articles based on their future number of comments [66]. Using a data set of news articles and comments, the authors showed that a linear-log method could be

an effective solution for automatic online news ranking.

Predicting the popularity of web items, based on the aggregate user behavior, has also been addressed as classification problem. Jamali and Rangwala used the number of comments that Digg stories receive in the first ten hours to predict the final Digg score [49]. By training different classification methods (decision tree classifier, k-nearest neighbor, and SVM), they show that it is possible to predict the popularity class of a Digg story with an accuracy of 80%, 64%, and 45% when considering a separation in 2, 6, and 14 ranges of popularity. Hong et al. studied the problem of predicting the number of retweets for Twitter posts [53]. The authors addressed this problem as a multi-class classification task, where, for a given tweet the goal is to predict the range of popularity and not the exact retweet count. Using a logistic regression classification function and various content, topological, and temporal features the authors showed that they can successfully predict which messages will not be retweeted (99% accuracy) and those which will be retweeted more than 10,000 times (98% accuracy).

Temporal analysis. For web content that captures users' attention for longer periods of time (e.g., certain videos that are requested during several months or even years) it has been observed that the aggregate-based prediction models are prone to large errors [43]. To improve the prediction effectiveness, one immediate solution is to design models that can weight users' attention differently based on the recency of the information relative to the prediction moment. For this type of evaluation, the aggregate user behavior is sampled in equal-size intervals of duration δ where $x_c(i)$ is the popularity of an item c during the ith interval, and $X_c(t_i)$ the vector of popularities for all intervals up to t_i: $X_c(t_i) = [x_c(1), x_c(2), x_c(3)..., x_c(i)]^T$ ($N_c(t_i) = \sum_{j=1}^{i} x_c(t_j)$).

This type of approach has been used by Pinto et al. to predict the popularity of YouTube videos [69]. Using a sampling rate of one day the authors propose the use of a `multivariate linear regression` expressed as

$$\widehat{N}_c(t_i, t_r) = \Theta(t_i, t_r) X_c(t_i) \cdot \tag{2.14}$$

The parameters of the model, $\Theta(t_i, t_r) = [\theta_1, \theta_2, .., \theta_i]$ are computed to minimize MRSE under the new definition of estimated popularity. The performance of this model has been empirically studied on a collection of YouTube videos and showed a significant improvement compared to `constant scaling` model (an aggregate-based prediction model). For instance, predicting the popularity of a video one-month after publication using data from the first week showed an average improvement of 14% over the constant scaling model. The main drawback of this algorithm as stated by the authors, is that, in order for the prediction methods to be effective, additional exploration is needed to decide on the optimal history length and the sampling rate.

`Reservoir computing` [70], a novel paradigm in recurrent neural networks, has been used to study more complex interactions between early and late popularity values (between $X_c(t_i)$ and $N_c(t_r)$). More specifically, this technique is used to build a large recurrent neural network that allows one to create and evaluate nonlinear relationships between $X_c(t_i)$ and $N_c(t_r)$ [71]. The model was tested on a small sample of YouTube videos and showed a minor improvement over `constant scaling` model in predicting the daily number of views based on the observations received in the previous ten days.

For videos that are popular over very long periods of time (those that receive views during at least half a year), Gursun et al. [29] observed that daily view counts can be modeled through a `time series prediction` model using Autoregressive Moving Average (ARMA). The popularity of a video at a given day n, $x_c(n)$, can be computed with the following formula:

$$x_c(n) = \sum_{i=1}^{p} \alpha_i x_c(n-i) + \epsilon_n + \sum_{j=1}^{q} \theta_j \epsilon_{n-j}, \tag{2.15}$$

where $\alpha_1, ..., \alpha_p$ are the parameters of the Autoregressive model, $\theta_1, ..., \theta_q$ are the parameters of the Moving Average, and $\epsilon_n, \epsilon_{n-1}, ...$ are the white noise error terms.

The model showed good performance in predicting the number of daily views using observations received in the previous week ($p = q = 7$), with an average error (MRE) of 15%. This result suggests that using the number of views received during one week is sufficient to predict the number of views during the next day. The main limitation of this method is that it has a very high computational cost as it requires one ARMA model per video. To improve the scalability of the model the authors use principal component analysis (PCA) to reduce the number of ARMA functions: use PCA to find the main principal components that can approximate the time series for the entire collection of videos and apply ARMA modeling to the principal components instead of the individual time series. This solution significantly improves the scalability (e.g., it requires 20 ARMA models for the entire collection of videos) with minor decrease in the prediction accuracy (MRE = 0.12 when using individual ARMA models compared to MRE = 0.14 when using principal component analysis).

Kong et al. proposed kSAIT (top-k Similar Author-Identical historic Tweets), an algorithm to predict the popularity of tweets 1, 2, or 3 days after publication based on the retweet information received in the first hour [55]. The underlying assumption of this algorithm is that, tweets are retweeted in a similar manner depending on the author of the tweet. The prediction algorithm is thus user-specific (there is one prediction function for each user) and uses as predictive features only users' retweeting behavior (it does not include any information about content itself or about users' centrality in the graph of social

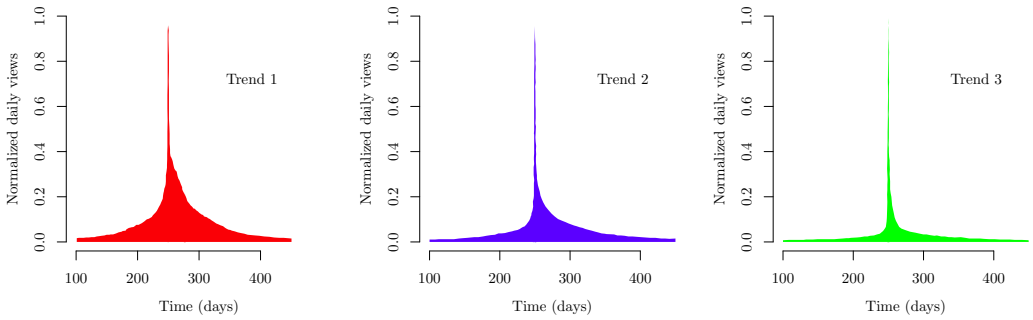


Figure 2.3: Example of three popularity evolution trends discovered by Crane and Sornette [35] (and similar with some of the trends presented by Figueiredo [36] and Gursun et al. [29]). The figure shows the average number of views centered and normalized by the popularity during the peak day.

interactions).

Each tweet is described by a set of features (e.g., retweet acceleration, retweet depth) derived from the time-series of retweets published in the first hour after publication by the direct and n-level followers, the publication time of the tweet, and information about the users who retweet the original tweet. When a new tweet is posted, the algorithm computes the similarity of the tweet and all other tweets published by the same user, selects the top-k most similar tweets, and estimates the popularity of the target tweet as an average of the popularity of the top-k most similar tweets.

The performance of the algorithm has been studied on a data set from Tencent Weibo and compared to several regression-based methods. The algorithm showed good prediction performance (and improvement of up to 10% in terms of MAE compared to regression-based methods), but training a personalized function for each user makes it difficult to implement in large-scale social networks.

Popularity evolution trends. Several studies showed that the popularity of different web items follows a similar evolution over time [29, 35, 36, 72]. For example, Crane and Sornette have revealed three common trends (illustrated in Figure 2.3) in the evolution of daily YouTube views that correspond to viral, quality, and junk videos [35]. Detecting these trends is useful as they provide a richer information about the evolution of content popularity that could further be exploited in the prediction process.

For rarely-accessed videos (those that are viewed less than half a year), Gursun et al. observed that most videos experience similar popularity evolution patterns around the peak period [29]. To reveal these patterns the authors employ `hierarchical clustering` using the time-series of video popularity during 64 days centered around the peak. This

allows them to observe that there are ten common shapes that describe the evolution of popularity for most of the videos. After these shapes have been detected, the prediction task consists in mapping videos to the clusters that best describe their evolution until the prediction moment (t_i) and in using the temporal evolution trends of the clusters to deduce future video popularity. On a sample of YouTube videos, this method showed good performance in making short term predictions (predict the number of views in the next day) but significantly larger ones in seeing further.

Pinto et al. extended the multivariate linear regression model by including additional information that captures the similarity between videos in terms of their temporal evolution patterns [69]. The model assumes that the popularity evolution of a subset of videos ($C_1 \in C$) is representative for the entire population and could be used in the prediction process. More specifically, the prediction model, called `multivariate radial basis function` (MRBF), is described by the following relationship:

$$\widehat{N}_c(t_i, t_r) = \Theta(t_i, t_r) X_c(t_i) + \sum_{c_1 \in C_1} w_{c_1} \mathrm{RBF}_{c_1}(c), \tag{2.16}$$

where $C_1 \in C$ is the representative subset and w_{c_1} is the weight associated with each member. RBF is the Radial Basis Function with Gaussian kernel that measures the similarity between the target video and each video in C_1 [73, chapter 6]. Training MRBF involves finding the optimal parameters Θ and w_{c_1} to minimize MRSE, setting the optimal values of RBF kernel, and finding a representative set C_1. The performance of this model showed an average improvement of 5% over multivariate linear regression and 20% compared to the constant scaling model.

Ahmed et al. proposed a model that uses a more granular description of the temporal evolution of popularity [38]. Instead of using a set of representative web items to describe the entire evolution of content popularity, this model considers representative members for each interval δ_i and defines rules to model the transitions among subsequent intervals.

The representative members for each interval, acting as representatives for clusters of popularity, are computed using Affinity Propagation clustering algorithm [74]. To calculate the similarity between items, the authors derive two features from $X_c(t_i)$: one that compares if two items receive the same proportion of users' attention and another that measures if the two items experience a similar popularity growth. Once the clusters of popularity have been identified, they are grouped into a probabilistic framework used to describe the evolution of content popularity between clusters over time. Thus, by knowing to which cluster an individual item is most likely to belong at time t_i, one can predict its popularity at a future moment of time t_r.

The performance of the model has been tested on three data sets (YouTube, Vimeo,

and Digg) and showed a significant improvement over the log-linear model. For example, when using the observations received in the first 24 hours to predict the popularity 4 hours ahead, this model led to an error (MRSE) of 1% (Digg) and 3.5% (Vimeo and YouTube) compared to 17% (Digg), 24.2% (Vimeo), and 29.7% (YouTube) when using the log-linear model.

2.5.1.3 After publication - Individual behavior

Instead of treating each user reaction equally in the prediction process, models under this category draw conclusions based on individual user behavior.

`Social dynamics`, the model proposed by Lerman and Hogg, describes the temporal evolution of a web content popularity as a stochastic process of user behavior during a browsing session on a social media site [75]. In its original form, it was designed according to the characteristics of the social bookmarking site Digg: stories can be found in three sections of the site (front, upcoming, and friend list pages), users can express their opinions through votes, and stories are arranged in pages (or promoted to different sections of the site) based on the dynamics of votes.

User behavior is modeled through a set of states that describe the possible actions that one can take on a site: browse through the different sections, read news stories, and cast votes to further recommend them to the Digg community. Browsing sessions are dynamic as stories circulate through the site (e.g., they may appear on different sections of the site or change position on the page) depending on the voting results. Individual user behavior is thus linked to the collective behavior, which in the end explains how stories receive votes over time. More specifically, the number of votes a story receives depends on its visibility and general interest. Visibility is expressed as the probability of finding a story in different sections of the site and the interest is linked to the quality of the story estimated by the voting dynamics.

The authors validated the model on a small sample of Digg stories by studying user reactions to the publication of stories and by taking into account the relationships between Digg users. By using this algorithm, the authors show that they can predict in 95% of the cases which stories will become popular enough to reach Digg's front page. In terms of numeric prediction of the number of votes, results indicate that the first twenty votes are strong predictors of the final Digg score (RMSE = 593, compared to 610 when using log-linear model).

For platforms that allow users to cast positive and negative votes, Yin et al. proposed `Conformer Maverick` model used to predict web content popularity based on user voting profile [60]. The underlying assumption of the model is that, in the voting process, users can have two behaviors: obey the general user opinion (the "conformers") or be against

them (the "mavericks"). User personality (its profile) is in-between these two extremes but in general one trait prevails.

The first step is to build user profiles based on the voting history by comparing individual votes with the overall appreciation of the content (if the majority of votes is positive or negative). These profiles are later used to decide if an item will be popular by analyzing early user votes. Receiving positive votes from conformers and negative ones from mavericks is then a good indication that one piece of content will be appreciated by the majority. The algorithm has been tested on a joke sharing application (JokeBox) and showed better performance than collaborative filtering solutions.

Zaman et al. propose a probabilistic model based on Bayesian inference to predict the popularity of Twitter messages [56]. The predictive features are content-agnostic and based on retweets time-series and the social connectivity graph of the Twitter users. The model is based on the assumption that Twitter users have similar actions with regard to the post of a tweet (to share or not) that creates a pattern in the evolution of tweets popularity. More particularly the probability of a (re)tweet to be retweeted depends on the number of followers and the distance from the user that originally generated the tweet. Using a small data set of 52 tweets, the method showed a good performance (given the difficulty of task), with an average error (MRE) of 40% using the retweeting information received in the first 5 minutes after the publication.

2.5.2 Cross domain

The second major category of methods are used to predict web content popularity using information from multiple web domains: they extract data from one domain (e.g., social media) and transform it into knowledge used to predict web content popularity in another domain (e.g., the site where content has been published). Currently, only methods that predict content popularity *after publication* based on the *aggregate behavior* have been designed.

Oghina et al. used information from Twitter and YouTube to predict movie ratings on IMDb [58]. By training a linear regression model on several textual features extracted from Twitter and various statistics from YouTube (likes, dislikes, and comments) the authors showed that they can accurately predict movie ratings on IMDb. The most accurate prediction model has proved to be the one that combines the ratio of likes over dislikes from YouTube activity with the subjective terms (positive and negative unigrams about the movies) extracted from Twitter.

The algorithm proposed by Roy et al., `Social Transfer`, extracts information from Twitter to detect videos that will experience sudden bursts of popularity on YouTube [76]. The model can be decomposed in the following three steps: extract popular topics from

Twitter, associate these topics to YouTube videos, and compare the popularity of videos on Twitter with their popularity on YouTube. A disproportionate share of attention on Twitter compared to YouTube is then used as strong evidence that an individual item will experience a sudden burst in popularity.

Topics are learned by analyzing Twitter stream, extracting topical words, and finding topics from words with semantic similarity. Each topic has a certain popularity on Twitter (called social prominence) based on its prevalence in the Twitter stream and the time it first appeared. The algorithm uses the Social Transfer framework [77] to map videos – using only the textual information from the title and video description – to topics extracted from Twitter. The popularity of a video on Twitter (expressed by the popularity of its topic) is then compared to its popularity on YouTube (represented by number of views) and, if the difference is significant, the video is susceptible to receive a sudden burst of attention.

Using data from YouTube and Twitter, and by training a support vector machine (SVM) classifier, the algorithm showed that it can predict with 70% accuracy videos that will experience a significant increase in popularity on a daily basis. The results show an improvement of almost 60% compared to a model that uses only the information available on YouTube.

Another approach used in cross domain content popularity prediction is the method proposed by Castillo et al. that collects information about the early attention that news articles receive on social networks to predict the total number of page views on a news site [19]. The statistical method used for this task is a `multiple linear regression` that uses as input variables: number of Facebook shares, number of tweets and retweets, entropy of tweet vocabulary, and mean number of followers sharing the articles on Twitter. Using a collection of Al Jazeera news stories, the authors showed that a model based on the social media signals received in the first ten minutes after publication achieves the same performance as one based on the number of page views received in the first three hours.

These results show that, when information related to a web content is spread across multiple web domains, aggregating information from multiple sources can significantly improve the prediction accuracy. In particular, the information extracted from Twitter proved very useful in learning more accurate prediction models. The benefit of using social streams as an additional source of information can be explained by the fact that sharing (and resharing) is one of the most popular methods to reach information on the Internet. And, as sharing rarely happens inside the originating web domain, this information gives an additional – and more reactive – perspective about the popularity of a web content.

Table 2.1: Summary of the popularity prediction methods.

Class	Methods	Data set	Comparison	Performance / Remarks
Before publication	SVM, Naive Bayes, Bagging, Decission Trees, Regression [65]	Feedzilla		An accuracy of 84% in predicting the popularity range of an article.
Before publication	Random Forests [64]	AD,De Pers, FD, NUjiji, Spits, Telegraaf, Trpuw,WMR		Good performance in identifying articles that will receive at least one comment.
Cumulative growth	Constant growth [46]	Slashdot		Good performance in predicting the volume of comments one day after the publication of an article (MSE = 36%).
Cumulative growth	Constant scaling [43]	Digg, YouTube	Constant growth, Log-linear	Outperforms the other two methods in terms of MRSE.
Cumulative growth	Log-linear [43]	Digg, YouTube	Constant growth, Constant scaling	Outperforms the other two methods in terms of MSE.
Cumulative growth	Survival analysis [57]	DPreview, MySpace		Using the information received in the first day it can detect with an 80% accuracy threads that will receive more than 100 comments.

Continued on next page

Table 2.1 – continued from previous page

Class	Methods	Data set	Comparison	Performance / Remarks
Cumulative growth	Logistic regression [53]	Twitter		Successfully predict which messages will not be retweeted (99% accuracy) and those who will be retweeted more than 10,000 times (98% accuracy).
Temporal analysis	Multivariate linear regression [69]	YouTube	Constant scaling	An average improvement of 15% (in terms if MRSE) compared to constant scaling model.
Temporal analysis	Reservoir computing [71]	YouTube	Constant scaling	Minor improvement over constant scaling model.
Temporal analysis	Time series prediction [29]	YouTube		For frequently-accessed videos. Good performance in predicting daily views.
Temporal analysis	kSAIT [55]	Twitter	Regression-based methods	Predict the number of tweets using information from the first hour. An improvement of up to 10% compared to regression-based methods.
Popularity evolution patterns	Hierarchical clustering [29]	YouTube		Designed for rarely-accessed videos. Good performance for short-term predictions but significantly larger ones for long-term predictions.
Popularity evolution patterns	MRBF [69]	YouTube	Constant scaling, Multivariate linear regression	An average improvement of 5% (in terms if MRSE) compared to multivariate linear regression and 21% compared to constant scaling model.

Continued on next page

Table 2.1 – continued from previous page

Class	Methods	Data set	Comparison	Performance / Remarks
Popularity evolution patterns	Temporal-evolution prediction [38]	YouTube, Vimeo, Digg	Log-linear	Significant improvement compared to log-linear method. It can be used to predict the temporal evolution of popularity.
Individual behavior	Social dynamics [75]	Digg	Log-linear	It incorporates information about site's design. It can be used to predict the temporal evolution of popularity.
Individual behavior	Conformer Maverick [60]	JokeBox	Collaborative filtering solutions	Adequate for platforms that rank content based on user votes.
Individual behavior	Bayessian networks [56]	Twitter		Predict the total number of tweets using the information received in the first 5 minutes after publication. MRE = 40%
Cross-domain	Linear regression [58]	IMDb, Twitter, YouTube		Predict movie ratings using social media signals. The best performance was achieved when using textual features from Twitter and the fraction of likes over dislikes from YouTube.
Cross-domain	Linear regression [19]	Al Jazeera		A model based on social media reactions in the first ten minutes has the same performance as one based on the number of views received in the first three hours.

Continued on next page

Table 2.1 – continued from previous page

Class	Methods	Data set	Comparison	Performance / Remarks
Cross-domain	Social dynamics [76]	YouTube, Twitter	SVM basic	70% accuracy in identifying videos with sudden bursts in popularity (60% improvement over a model that uses only the information available on YouTube).

2.6 Selecting the right features

The quality of the prediction is reflected by the quality of the data used as input in the prediction model. While the majority of the models presented earlier make popularity estimations based on early popularity counts (e.g., number of views/comments received in the first hours after publication to predict the number of views/comments later in time) some of the models rely on richer information in the prediction process. We make a brief summary of the various features used in the prediction models and report their predictive capacity.

Characteristics of content creators. The online media ecosystem is populated by content creators – independent producers, professional bloggers, mainstream mass media, or news agencies – with different but relatively stable audience that could be used as additional knowledge in the prediction process. This information has been exploited by Bandari et al. who showed that, when predicting the number of tweets that an article will generate, one of the strongest predictor is the publisher of the news article [65].

Content features. It has been observed that certain words or key phrases, that probably refer to hot or controversial topics, often produce a significant amount of attention. Tsagkias et al. observed that the top most popular terms used in the text of news stories have a strong and robust performance in predicting if articles will receive comments and if the volume of comments will be high [64]. Similar, Marujo et al. observed that the highest prediction performance (when predicting the popularity of news articles) was obtained after including in the prediction model popular key-phrases from the text of the articles [78].

Content category. Designing specialized prediction models for the different content categories showed little benefit in predicting the popularity of videos [69] and new articles [65,78]. The only notable exceptions have been signaled for YouTube *Music* videos [69] and news articles related to *Technology* [65]. The low predictive performance of using this

information in a prediction model can be explained by the overlapping scope of categories, with content often belonging to multiple categories [32, 65].

Named entity identification. The popularity of people, locations, or organizations may be directly correlated to the popularity of the web item they are associated with. Tsagkias et al. found that including popular entities from Netherlands in a prediction model is useful in spotting articles that will be commented [64].

Sentiment analysis. Text is often charged with emotion that may be appealing to online readers. The subjectivity of the language has shown little predictive power in predicting the volume of tweets for online news stories [65]. However, it has been observed that articles that are written in a more positive or negative voice, associated with strong emotions (e.g., admiration or anger), are good indicators of how viral the article will become [79]. In addition, Oghina et al. observed that extracting subjective terms from the discussions about movies on Twitter can be used to build regression models that can predict movie ratings on IMDb [58].

Comment statistics. Jamali et al. used various comment statistics (number of comments, average word length, and the hierarchical organization of comments) to predict the popularity of Digg stories and found that the number of comments is a strong predictor for the Digg score [49].

Social media signals. As we saw in Section 2.5.2, social media conveys valuable information about the popularity of a web content. Castillo et al. showed that the attention that news articles generate across social networks (number of Facebook shares, number of tweets and retweets, the language of the Twitter messages) is effective in predicting the popularity of articles on a news site [19]. Another example of the predictive power of social media has been reported by Roy et al. who showed that the popularity of a topic on Twitter is a good indication that a YouTube video will experience a sudden burst in popularity [76].

Social sharing viewing behavior. Recent applications such as Yahoo! Zync allow users to share and jointly manipulate content in real time, which produces additional digital traces that can be used in content popularity prediction. Shamma et al. studied how users' actions during a sharing session can be used to predict the popularity of YouTube videos and observed that these interactions are strong indicators of videos with a high number of views [59].

Real-world features. Content published in online media is strongly related to real-world events but transferring information from physical to online world is difficult. An attempt to employ real-world information in the predictions process has been done by Tsagkias et al. who showed that there is an insignificant benefit in using weather conditions (average temperature in Netherlands) to predict the volume of news comments [64].

2.7 Factors that influence content popularity

Research on web content popularity has evolved from describing the popularity characteristics to understanding the temporal evolution as well as designing models to predict the future. However, during this process, little has been said about the factors that can drive a web content to its success. We report the main findings on the factors known to have an important impact on the popularity growth.

The amount of attention that a web item generates depends on various content and content-agnostic factors. In general, the content itself explains much of its popularity. Creating quality content [80, 81], that generates strong emotions [79], and has a large geographic relevance [82, 83] is more likely to attract a larger audience. The topic of the content is also important, as popularity is susceptible to bursts of attention in response to real-world events [84]. On the other hand, there are elements that have a negative impact on content popularity. One of them is the presence of multiple versions of the same content that tends to limit the popularity of each individual copy [32].

There are also several content-agnostic factors that have a strong impact on the popularity growth [85]. Popular Internet services, such as search tools, recommendation systems, and social sharing applications can extent the visibility of a web item and increase its popularity. Taking the example of YouTube (one of the most active platform for this kind of studies) the internal search engine accounts for most of the views, followed by the recommendation systems, and the social sharing tools [17, 85]. But the outcome of these services can also play an important role in the popularity outcome. For example, it has been observed that videos have a higher chance of becoming popular if they are placed in the related list of popular videos [20, 86] and higher the position of the video in the list the greater the number of views [87]. The recommendation system thus creates a strong linked structure between similar videos, which influence each other in terms of popularity [88]. This information can be extremely valuable to newborn videos that can have a big advantage in creating relationships with similar popular videos by choosing a relevant title, description, or keyword set [88].

Social sharing acts as an additional catalyst of user attention. Diffusing videos though social networks, blogs, or e-mail services generates peaks of attention during short periods of time but the popularity quickly drops afterwards [63]. Similarly, the social interactions created within a site play an important role in the success of a web object. This factor is particularly important in the early stages of the objects's lifetime for which the bigger the social network of the publisher the greater the increase in popularity [85]. Finally, social influence can have a non negligible consequence in the popularity growth. A study conducted by Salganik et al. has revealed that, when users were informed about the

collective decisions of other individuals, the popularity of songs were driven by a "rich-get-richer" process where early user attention explained much of the later one [80].

2.8 Predictive proactive seeding: an application of web content popularity prediction

The ability to predict future web content demand can prove valuable to different actors: online users can filter more easily the massive amount of information; content producers and content providers can better organize their information and build more effective delivery platforms; and advertising networks can design more sophisticated and profitable advertising algorithms.

In this dissertation we study one concrete application that can benefit from popularity prediction methods: predictive proactive seeding in the context of mobile data offloading. To achieve this, in the following chapters we will first study the capacity to predict the popularity of a specific type of web content (Chapter 3) and then analyze the possible gain that prediction methods can bring to a proactive seeding approach (Chapter 4).

2.9 Conclusions

In this chapter we reviewed the current state-of-the-art in online content popularity prediction. To structure the existing prediction methods we have proposed a classification based on the type of information used in the prediction process. We presented the different prediction methods, reported their performance, and looked at the different features that have been observed to have a strong predictive power.

Chapter 3

Predicting the popularity of online news articles

3.1 Introduction

To better understand the problem of predicting the popularity of web content popularity, given the panoply of methods applied to different types of web items, we study the feasibility of predicting the popularity of web content using real web traces. We chose as content of interest online news articles, an engaging type of online content that captures the attention of a significant amount of Internet users. This is a type of content that can easily be produced, has a small size, and low cost – properties that makes it interesting to be massively spread through online social platforms and particularly enjoyed by mobile users.

We analyze two important online news platforms from France and Netherlands, provide insights on how articles are produced and consumed by online readers, and study the effectiveness of predicting their popularity. We focus on one dimension of the content popularity and consider the *number of comments* as an implicit evaluator of the interest generated by an article. As we saw in Chapter 2, predicting the popularity of online content is a complex and difficult task and different prediction methods and strategies have been proposed in the literature. We select as case studies two of these methods, that are adequate to the type of information contained in our data set, and study their capacity to predict the popularity of online news.

In addition to predicting the exact amount of attention that one content will generate, in another practical situation, it may be valuable to rank articles based on their future popularity. As the online users' interest in web content is often highly skewed (with

popular objects being extremely popular) finding the top most popular objects is often a good-enough solution for applications that benefit from predicting users' preferences. For example, a *Top-10* approach to prefetching content has proved to be robust solution to anticipate future content requests [89]. Thus, we study the effectiveness of using prediction methods to rank news articles based on their future popularity and compare them with various heuristics and more customized learning to rank algorithms.

3.2 Background

In Section 2 we described the various methods used to predict the popularity of web items and showed that different approaches have been proposed depending on the type of information used in the prediction process. We saw that the choice of a method (and its accuracy) depends on the type of web item, the granularity of user behavior (using aggregate of individual user information), or the possibility to cross-correlate information from multiple web sources. In this chapter we study the possibility to predict the popularity of news articles using the aggregate user behavior.

We position this work with respect to similar studies that address the problem of predicting the popularity and ranking online news articles. One of the first models, used to predict the popularity of Slashdot stories, was proposed by Kaltenbrunner et al. [90]. This solution considers that, depending on the publication hour, the popularity of news stories follows a constant growth. Szabo et al. proposed two other prediction methods that have shown good results in predicting the popularity of Digg stories [91]. Tsagkias et al. showed that a linear model on a logarithmic scale (used in [91]) is also reliable method for predicting the popularity of news articles [92]. Lerman et al. propose a model based on the social influence and web platform characteristics in the prediction process [93]. A different approach was proposed by Lee et al. where, instead of predicting the exact popularity value, the authors are interested in predicting the probability that a content will continue to receive comments after a certain period of time [94]. We place ourselves in this context of popularity of online news stories. In our work we analyze the capacity to predict the popularity of online news articles using methods that are adapted to our prediction goal (predict the exact popularity count at a certain moment in time) and can operate with the the type of information available in our data sets (we lack information about the social influence and web site's characteristics). We make a step further in our research and analyze the ranking capabilities of these methods by taking into consideration the dynamic nature of news generation.

The feasibility of ranking online news has been addressed in [95, 96]. McCreadie et al. propose a ranking method based on relevant blog posts and show that the blogosphere

Table 3.1: Summary of the data sets analyzed in this paper.

Data set	20minutes	telegraaf
Lifespan:		
- start	3/2/2007	18/8/2008
- end	6/5/2011	21/4/2009
Total articles	231,120	40,287
Total comments	2,635,489	731,395
Articles per day		
- mean	157	176
- median	136	153
Comments per day		
- mean	1,255	3,086
- median	1,231	3,052

activity is a reliable indicator of news stories importance [96]. A different approach was proposed by Morales et al. who use a learning to rank algorithm and Twitter posts to rank news articles based on the future number of clicks [95]. The study shows that micro blogging activity can successfully be used to detect the important news stories. In our study we share the same general objective of ranking news articles, but our work differs both in the ranking technique, notion of article relevance, and input used for the ranking methods.

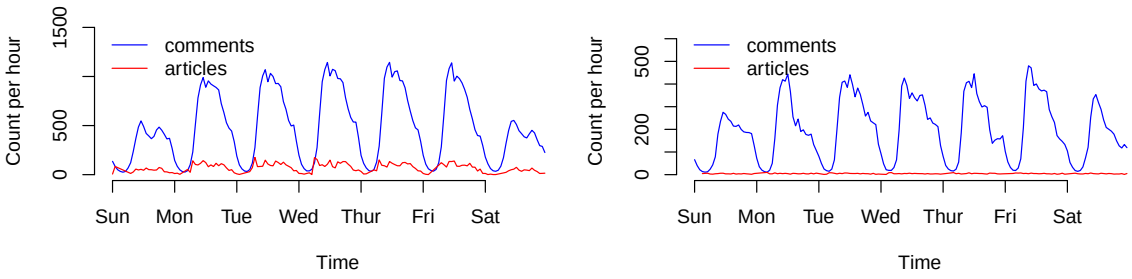
3.3 Global statistics

3.3.1 Online news data collections

In this study we use data from two news platforms, `20minutes`[1] and `telegraaf`[2]. Both news sources are popular daily newspapers that complement the hard copy editions with online sites that allow users to read news stories and express their opinions through comments. The sites' content is news oriented, starting with the main articles from the printed version and being periodically updated with the latest news. These newspapers target a broad audience and cover diverse topics from national and international politics, sports, economy, or lifestyle.

The two data collections differ in size and lifespan: `20minutes` contains 231,120 articles and 2,635,489 comments published from February 2007 until May 2011 [97]; `telegraaf` data set contains 40,287 articles and 731,395 comments published from August 2008 until April 2009 [92, 98]. We present a summary of the data sets in Table 3.1.

[1] http://www.20minutes.fr/
[2] http://www.telegraaf.nl/



(a) 20minutes (b) telegraaf

Figure 3.1: The number of articles and comments posted per hourly cycles. On the y-axis we illustrate the average number of articles and comments published per hour.

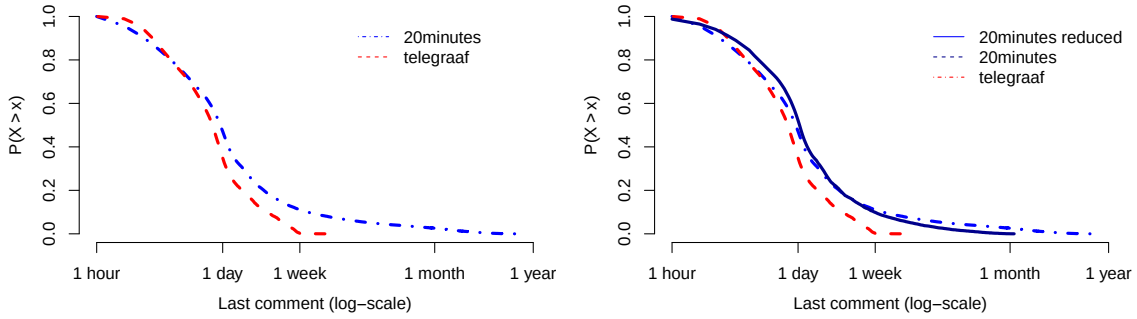
Figure 3.1 shows the average number of articles and comments published on an hourly basis during the period of one week. Our data sets confirm previous observations of the circadian pattern of user activity [40, 42]. Daily variations can also be deducted from this graph. Readers are twice more active during the working days compared to the weekend, nevertheless with an important contribution if we consider the number of published articles (fewer during weekends).

3.3.2 News articles lifetime

A common characteristic of online content is that it suffers from a decay of interest over time and, depending on the type of content, this interest is steep or gradual. News articles depict a very steep decay compared with videos [99] or photos [100] as they refer to a recent type of information that by its nature has a very short life cycle.

We provide a coarse representation of articles' lifetime by analyzing when articles received their last comment[3]. We present the results in Figure 3.2 by means of a complementary cumulative distribution function of the last comment time relative to articles publication time. For both news sources we observe that most articles, 72% for `telegraaf` and 61% for `20minutes`, receive all the comments within the first day after the publication. There are however articles that continue to receive comments after one day but in most of the cases they represent only a sparse interest and not a constant one as observed for other type of online content [99]. By analyzing the overall comments arrival rate, we observe that

[3]We are aware that there are other fine-grained methods of evaluating the decay of attention over time [94, 101, 102], but for the scope of our work, this coarse characterization provides us with sufficient information.



(a) (b)

Figure 3.2: Complementary cumulative distribution function corresponding to the articles' lifetime (time elapsed between article publication time and the last comment time). The labels on the x-axis correspond to one hour, day, week, month, and year. We represent two versions of `20minutes` data: one over the entire data set and a reduced version that covers the same period of time as `telegraaf` data set.

the most significant share of comments is received within the first day after the publication of an article. We report this observation in Figure 3.3 by means of a probability density function of the comments publication time relative to articles publication time. As it can be observed, for both news sources, users react very fast to news articles but their interest drops quickly after 6 hours and only a negligible amount of comments are received after one day.

Comparing the two news sources, we observe that, while the drop of interest over time is similar in the first day for both sites, articles published on `20minutes` engage users in a commenting activity for a longer period of time than those published on `telegraaf`. This difference can be explained by the different lifespan of the data sets, one covering more than four years and the other one only eight months. To isolate this effect we analyze a reduced version of the `20minutes` data set, one that covers the same period of time as `telegraaf` (Figure 3.2(b)). Even after this adjustment we can observe that, in general, `20minutes` articles receive comments for a longer period of time than `telegraaf`. There are several factors that could explain this difference. One of them is that `20minutes` news have a greater exposure than `telegraaf` news, as indicated by the traffic statistics of the two web sites (5.5 million unique visitors per month for 20minutes.fr compared to 3.8 million for telegraaf.nl[4]). The result is that `20minutes` articles may seize a greater amount of attention in the early stages after the publication, which could further impact the popularity and smoothen the decay of interest over time. Other explanations, which unfortunately cannot

[4]According to the latest statistics of the two sites: http://corporate.tmg.nl/en/result-second-quarter-2012 (`telegraaf`); http://www.mediametrie.fr (`20minutes`)

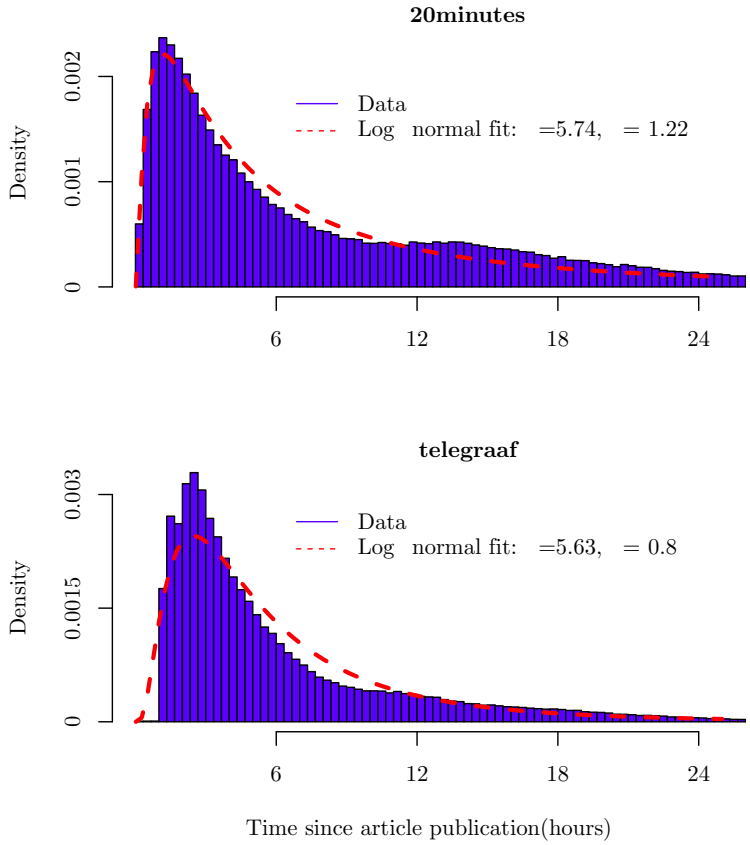


Figure 3.3: Probability distribution function of the comments time relative to the articles publication time. We represent the histogram covering a one day period along with the best probability fit, which in our case is best described by a log-normal distribution.

be deduced from the information found in our data sets, could be related to the tone of the articles (a more personal and subjective voice may be more captivating to online readers) or the topic of the news (it has been observed that certain topics have a longer life cycle [103]).

3.3.3 Distribution of popularity

A common question addressed by scientists that study the properties of online content is whether the data under observation exhibits heavy-tail characteristics or not. While this is interesting from a scientific point of view, where a mathematical model can summarize empirical data, this observation also has practical implications. For example, it has been shown that understanding the underlying distribution of popularity for web content can have important consequences in the design of caching algorithms [104, 105] or to improve

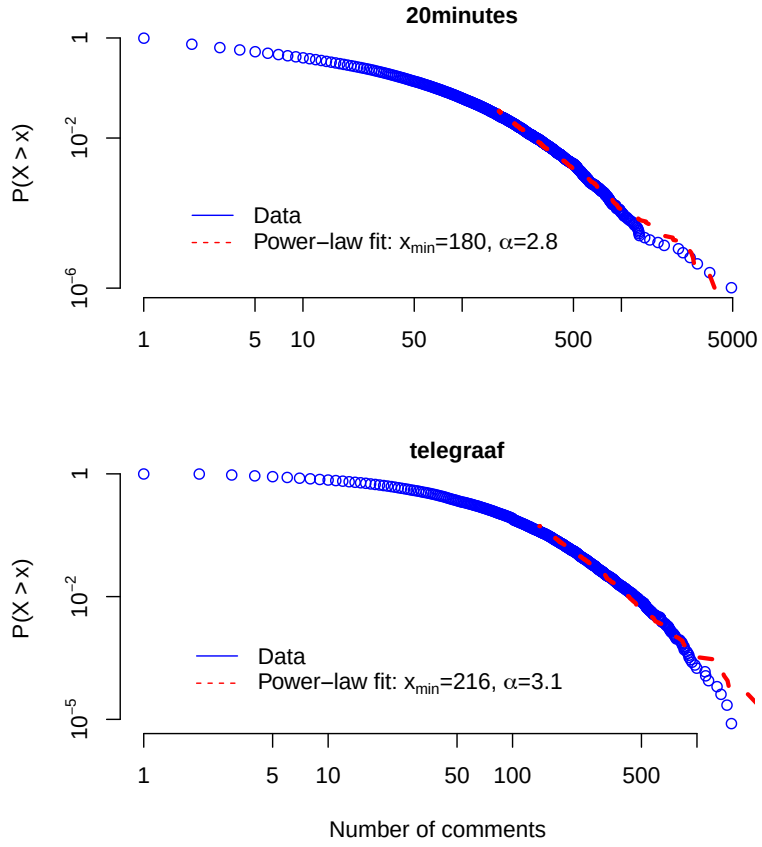


Figure 3.4: The complementary cumulative distribution function of the articles' popularity and the corresponding power-law fit.

the performance of search engines [106].

In the case of social media content, recent work, on different sources of online content and using various popularity metrics, indicates that content popularity can be described by heavy-tail distributions and the log-normal distribution appears to give the most consistent description [52, 98, 102]. Our data sets make no exceptions from this observation. This can visually be observed in Figure 3.4, where we present the complementary cumulative distribution of the number of comments per article and the power-law fit. The power-law appears in the tail of the distribution and has been confirmed by rigorous power-law tests proposed by Clauset et al. [107].[5] There is, however, a difference between the two news sources as observed in Table 3.2. Our results indicate that while a power-law is the most accurate solution for `20minutes` articles, a power-law with exponential cut-off is a better

[5]Statistical techniques based on maximum-likelihood methods and Kolmogorov-Smirnov statistics.

Table 3.2: Comparing the power-law fit against other alternative distributions. For each alternative distribution, we provide the *p-value* and the likelihood ratio test (LR). We consider a significance level of 0.1 for the *p-value* and display the significant values in bold. Positive values of the log-likelihood indicate that the power-law is a better fit model than the alternative distributions.

Data set	Exponential		Power + cut-off		Log-normal	
	LR	p	LR	p	LR	p
20minutes	34.42	**0.07**	-1.24	0.11	-2.5	0.31
telegraaf	13.40	0.12	-5.6	**0.00**	-4.6	**0.05**

alternative for `telegraaf` data.

It is out the scope of this work to debate over which distribution is the most adequate one for describing the popularity of online news and we encourage the reader to follow the enriching discussion presented in [108]. One possible explanation of why power-law is more visible for `20minutes` articles is given by the web site recommendation strategy. The site highlights the most commented articles in a dedicated section and twice a day it delivers to its subscribers a short electronic edition with the most commented articles. This creates a *rich-get-richer* effect, which is one of the reasons why power-law appears so often on the Internet [109]. The recommendation mechanism can also explain why the power-law fails to appear in the beginning of the distribution and could also account for the difference in articles lifetime observed in Section 3.3.2. Articles that are unpopular in the beginning do not benefit from any recommendation mechanism and the probability of receiving any kind of attention drops even more as they loose their position on the web site [101].

The heavy-tail property has important implications in the ranking evaluation. Indeed, given that the distribution is so heavily skewed, a ranking algorithm should perform particularly well in identifying the top most important articles. We explore in Figure 3.5 the exact spread of daily comments for the top most important articles. On the x-axis we order articles based on their popularity (in a decreasing way) and normalize the ranks from 0 to 100. For the y-axis we consider the proportion of the total comments (per day) received by the top-k most important. As we can observe in Figure 3.5, for both data sets, on a daily basis the top 10% most commented articles gather 50% of the total number of comments and around 20% of the articles receive 80% of all the comments published that day.

3.4 Predicting the popularity of online news articles

3.4.1 Popularity predictions methods

We consider the following two methods to predict the popularity of online news articles:

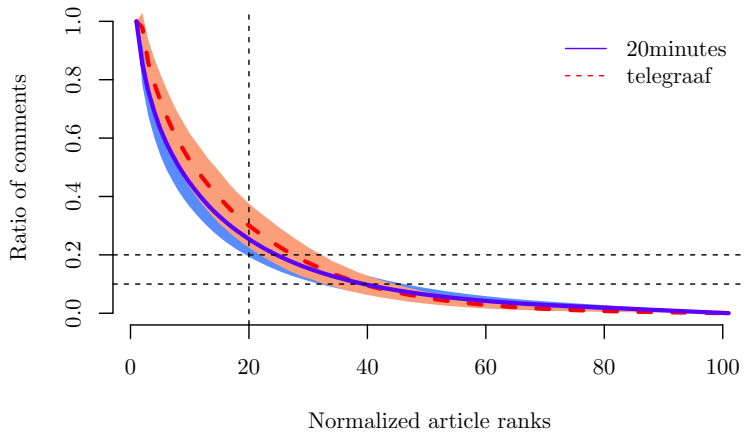


Figure 3.5: Normalized article ranks and the cumulative of proportion of comments received on a daily basis. We present the average value and one standard deviation (shaded area).

- Linear regression on a logarithmic scale model (**linear log**) proposed by Szabo and Huberman [91] and previously evaluated on Digg stories, YouTube videos, and Dutch news articles [98].
- **constant scaling** model also proposed by Szabo and Huberman and evaluated on Digg stories and YouTube videos [91].

The choice of the prediction model is given by the properties of our data, where the linear model on a logarithmic scale is particularly well adapted to data with heavy tail characteristics. We also consider the constant scaling model in our analysis following the observations that this model outperforms the *linear log* model with respect to the relative squared error [91].

These two models are regression functions where the dependent variable is the total number of comments an article receives at time t_r and the independent variable is the number of comments received t_i hours after the publication of an article. The goal of the prediction method is thus to estimate the number of comments t_r hours after an article a is published using the information received in the first t_i hours.

The estimated popularity for the *linear log* model is described by the following equation:

$$\widehat{N}_a^{\mathrm{LN}}(t_i, t_r) = \exp\left(\ln(N_a(t_i)) + \beta_0(t_i, t_r) + \frac{\sigma_0^2(t_i, t_r)}{2}\right). \tag{3.1}$$

For the parameters of Equation 3.1, β_0 is computed on the training set using maximum likelihood parameter estimation on the regression function $\ln N_a(t_r) = \beta_0(t_i, t_r) + \ln N_a(t_i)$ and σ_0^2 is the estimate of the variance of the residuals on a logarithmic scale.

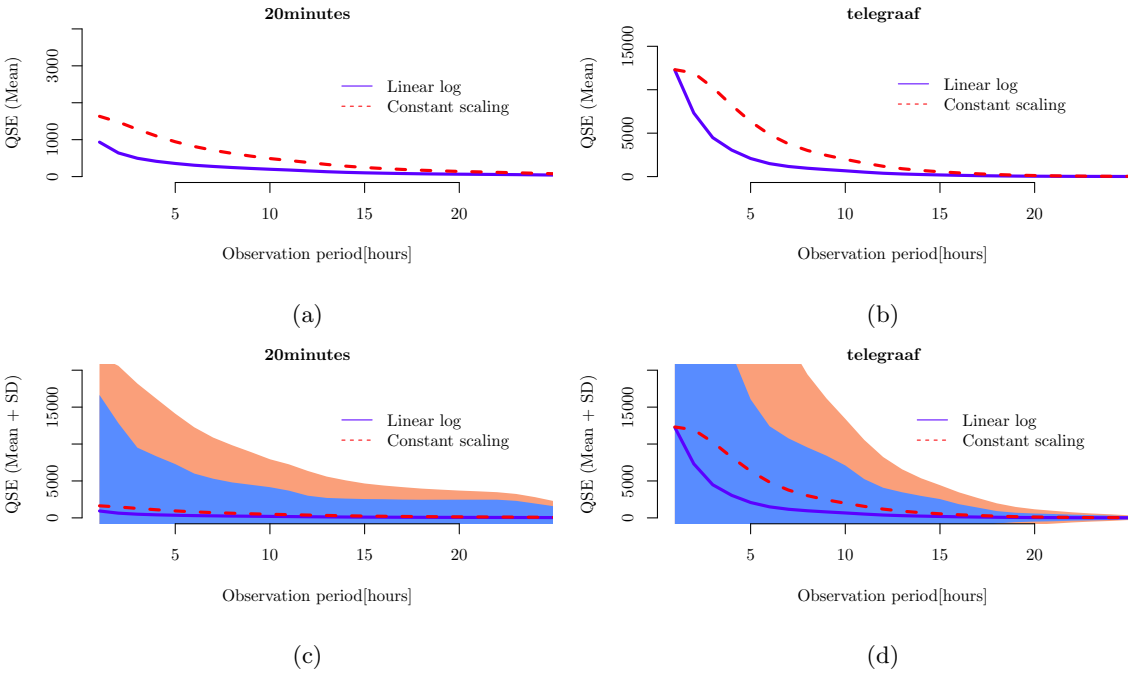


Figure 3.6: The prediction error in terms of QSE for the two popularity prediction methods. On the x-axis we vary the observation period from 1 to 24 hours. On the y-axis we represent the mean error (depicted in the top figures) and the mean along with one standard deviation represented by the shaded area in the bottom figures.

The *constant scaling* model is expressed as

$$\widehat{N}_a^{\mathrm{CS}}(t_i, t_r) = \alpha_2(t_i, t_r) \times N_a(t_i), \tag{3.2}$$

where α_2 is obtained using the following expression:

$$\alpha(t_i, t_r) = \frac{\sum_a \frac{N_a(t_i)}{N_a(t_r)}}{\sum_a \left[\frac{N_a(t_i)}{N_a(t_r)}\right]^2}. \tag{3.3}$$

3.4.2 Popularity prediction accuracy

We assess the performance of these methods in predicting the total number of comments using the absolute squared error (QSE) and the relative squared error (QRE):

$$\mathrm{QSE}(a, t_i, t_r) = \frac{1}{|A|} \sum_a [\widehat{N}_a(t_i, t_r) - N_a(t_r)]^2, \tag{3.4}$$

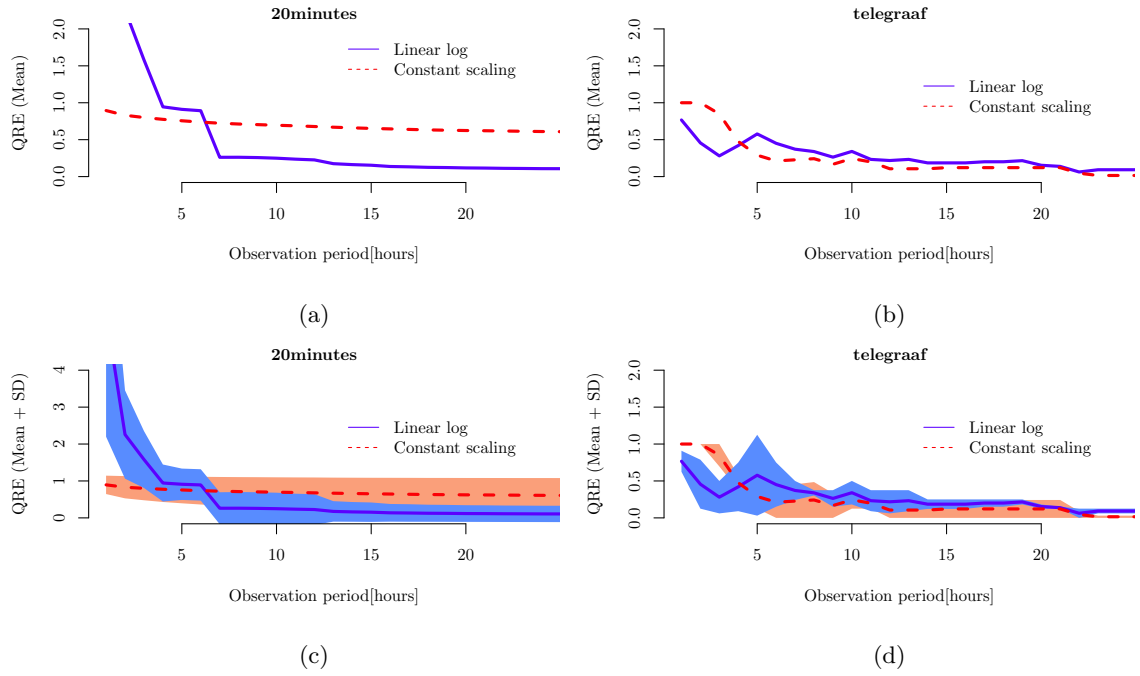


Figure 3.7: The prediction error in terms of QSE for the two popularity prediction methods. On the x-axis we vary the observation period from 1 to 24 hours. On the y-axis we represent the mean error (depicted in the top figures) and the mean along with one standard deviation represented by the shaded area in the bottom figures.

$$\mathrm{QRE}(a, t_i, t_r) = \frac{1}{|A|} \sum_{a} \left| \frac{\widehat{N}_a(t_i, t_r) - N_a(t_r)}{N_a(t_r)} \right|. \tag{3.5}$$

We analyze the predictive performance of these models as a function of the observation period (t_i) in Figures 3.6 and 3.7. The results indicate that the prediction error for both models is significantly high for an observation period of less than 6 hours and it rapidly decreases after that. Comparing the two data sets, we observe that `telegraaf` articles have very low predictive performance in the beginning and a negligible one after 20 hours. On the other hand, `20minutes` articles show a better overall predictive performance but the error prevails even after one day. The different performance of these models can, however, be explained by the different dynamics of the comment arrival rate presented in Section 3.3.2. As observed in Figure 3.1, the most significant share of comments is received in the first 6 hours after publication, which explains the high prediction error for short observation periods. Similar, the low error for `telegraaf` news stories after 20 hours is explained by the saturation of articles' popularity in less than one day.

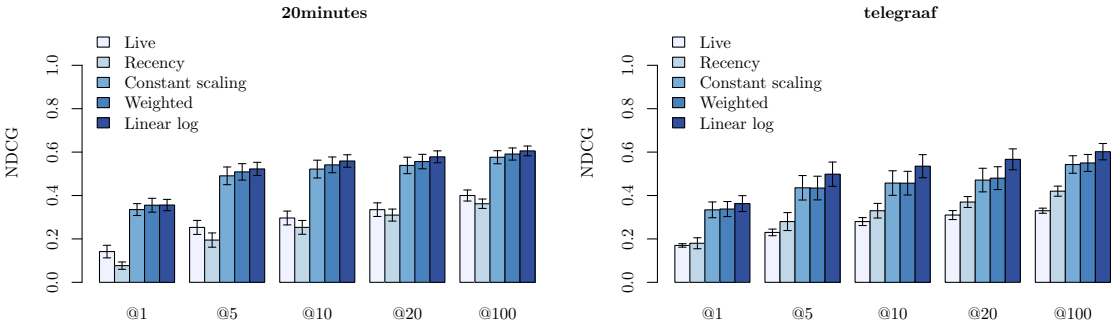


Figure 3.8: NDCG at different levels of precision. *@n* corresponds to the NDCG score for the top most important n articles. We present the mean over all prediction hours h (n=24) along with a 95% confidence interval.

3.5 Ranking news articles based on popularity prediction

Given that the popularity of web objects can often be described by Zipf-like distributions [104], the ability to detect only the most important ones is often a good enough solution in the design of applications that anticipate future clients' requests. For example, prefetching the top most popular documents from a server has been found as an effective prefetch heuristics [89] with a robust performance in anticipating future requests. In this case, an alternative to predicting the exact popularity value is to predict the relative order of the documents. We thus look at the ability of popularity prediction methods to identify the most popular articles and compare their performance with simple heuristics (that requires less computing effort) and dedicated learning to rank algorithms (that are more sophisticated techniques deigned for learning to rank problems [110]).

3.5.1 Methodology

To evaluate the ranking performance we propose the following methodology:

1. We break the corpus of articles of each data set in small subsets, where each subset contains all articles published during a certain *period* of time before a specific reference hour h. We set the duration of the *period* to one day given our previous observations of how readers significantly lose their interest in articles after one day.

2. We rank each subset of articles based on the number of comments that articles receive after the reference hour and consider this ranking as the ground truth. We then apply the different methods (heuristics, popularity prediction methods, and learning to rank algorithms) to estimate the ranking of articles and assess the ranking effectiveness using NDCG evaluation measure.

Ranking strategy. Let A be the corpus of articles published by a news platform during a period of time T, with $a \in A$ being one specific article. We discretize time on an hourly basis and consider h a precise hour of the day according to a 24-hour clock. Let t_h be the absolute time in hours and denote d as a one-day period. According to this time description and relative to an hour h we split A in k subsets, with $k = \lceil T/d \rceil$. Denote A_h^i the ith subset of articles created relative to an hour h, with $A = \bigcup_{i=1}^{k} A_h^i$. Please note that as h varies from 0 to 23 there are 24 ways of separating the corpus of articles. This separation allows us to further measure how the ranking performance is influenced by the hour we perform the ranking.

For every article a we refer to a_{t_0} as the article's publication time and define $N_a(t)$ the number of comments received by article a from a_{t_0} to certain time t. We also consider $N_a(t_h, t_r)$ the number of comments received by an article from t_h to t_r.

For this specific ranking task, given a set of articles A_h^i and a ranking time t_h, our goal is to accurately rank articles by the number of comments they will receive from t_h until a future time t_r, with $t_r > t_h$. We set t_r to 30 days to catch only the most relevant comments and reduce possible sources of spam. Under this description the ground truth ranking for A_h^i is given by $N_a(t_h, t_r)$. We consider this value the relevance of an article, and note

$$rel(a_{t_h,t_r}) = N_a(t_h, t_r)\cdot \tag{3.6}$$

Evaluation measure. We assess the ranking performance of the different strategies using the normal discounted cumulative grain (NDCG) [111]. To compute NDCG for a set of q articles we first determine DCG as

$$\mathrm{DCG} = rel_1 + \sum_{i=2}^{q} \frac{2^{rel_i} - 1}{\log_2(i+1)}, \tag{3.7}$$

where rel_i is the relevance of an article found at position i in the ranked list. From this value we compute NDCG as

$$\mathrm{NDCG} = \frac{\mathrm{DCG}}{\mathrm{IDCG}}, \tag{3.8}$$

where IDCG is the ideal DCG, the DCG of the perfectly ranked list of articles (ground truth ranking). We report the results using 10-fold cross-validation. That is, after splitting the corpus of articles in k subsets we randomly divide these subsets in 10 folds. We use 9 folds to train the models and assess their performance on the remaining fold; we repeat the process 10 times, using a different fold at each step, and report the average value.

3.5.2 Ranking methods

Each ranking method rates the relevance of an article using a certain criterion and one method is considered adequate if the estimated ranked list is close to the ground truth ranking. We analyze the ranking effectiveness of the two methods based on popularity prediction (*linear log* and *constant scaling*) and compare them with several baseline strategies:

- *Live*: rank articles by the number of comments received until the prediction moment, $N_a(t_h)$.
- *Recency*: rank articles by the time of publication, a_{t_0}, with the most recent first.
- *Weighted*: rank articles by the number of comments but weight the volume of comments per hour giving importance to more recent information.

The first two methods are simple heuristics often used by news portals to highlight their popular content, where *live* is oblivious to the temporal information and *recency* considers the time of the publication as the only factor that matters in the ranking decision. The third baseline method is similar[6] to the algorithm proposed by McCreadie et al. that showed one of the most accurate performance on TREC 2009 blog collection [96]. This method combines the partial popularity and recency of articles in the ranking decision by weighting the popularity relative to its closeness to t_h. By using this method, the score S assigned to an article a at time t_h is given by the following formula:

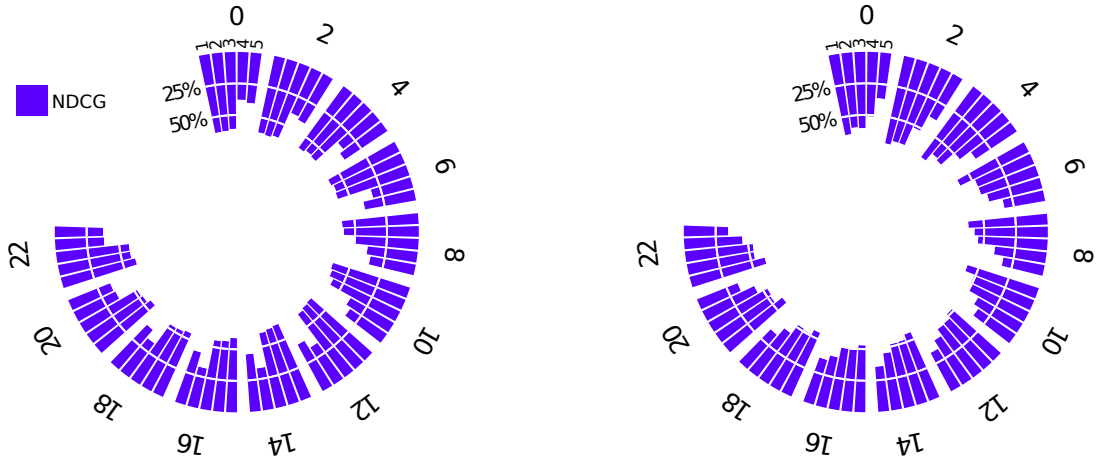
$$S(a_{t_h}) = \sum_{t=a_{t_0}}^{t_h} f(t_h - t) N_a(t) \cdot \tag{3.9}$$

where f is a probability density function that describes how much weight we should assign to past popularity on an hourly basis. In our case, we observed in Figure 3.3 that the decay of interest over time follows a log-normal behavior. As a result, we express f as log-normal probability density function:

$$f(\delta; \mu, \sigma) = \frac{1}{\delta\sigma\sqrt{2\pi}} \exp\left(\frac{(-\ln(\delta) - \mu)^2}{2\sigma^2}\right), \tag{3.10}$$

where we obtain the values of μ, σ by fitting the log-normal distribution on the empirical data.

[6]The algorithm uses the number of blog posts to predict users' interest in articles.



(a) **20minutes** (b) **telegraaf**

Figure 3.9: Ranking accuracy in terms of NDCG@100 per hourly basis. The outer numbers correspond to different reference hours h (only the even hours of the day). The inner numbers correspond to the different ranking methods, with 1 - linear log, 2 - weighted, 3 - constant scaling, 4 - recency, 5 - live.

3.5.3 Ranking performance

We compare the ranking performance of the two popularity prediction models with the baseline strategies (Figure 3.8). We report the mean value and a 95% confidence interval over all prediction hours and for various levels of precision: NDCG@1, NDCG@5, NDCG@10, NDCG@20, and NDCG@100. One can observe from the results that the simplest baseline models, *live* and *recency*, have limited ranking capabilities. This suggests that news ranking based on the submission time – *recency* heuristic – or one based on static view of the popularity – *live* heuristic – are inefficient solutions for this ranking task. The performance can however be improved using popularity prediction methods or a *weighted* solution. For a precision level of NDCG@100 (that allows us to capture on average 98% of the daily comments - Figure 4) the *linear log* model shows 50% improvement compared to *live* solution (for both data sets) and a 40% improvement for `telegraaf` - and 75% for `20minutes` - compared to the *recency* solution. From the top three performing algorithms, the *linear log* model shows the overall highest performance; the only exception is observed for NDCG@1, where the weighted model is equally effective. The gain of *linear log* model, compared to the second best solution (*weighted model*) for NDCG@100, is of 2% for `20minutes` and 10% for `telegraaf`. If the benefit brought by the *linear log* model over the other top two models is important for `telegraaf` (with an increase between 10% and 14% for precision levels greater than NDCG@5), for `20minutes` the top three methods show a similar performance

suggesting that they are equally fit for this ranking task.

These results depict the average performance over all hours of the day. However, in similar studies [91, 97, 98], it has been observed that articles and comments are published at a different rate during the day. As a consequence, articles may be more popular or exhaust their interest more quickly depending on the publication hour, an effect that can influence the ranking accuracy. To capture the impact of this observation, we illustrate in Figure 3.9 the ranking performance as a function of different prediction hours (to ease the presentation of the figure we report only the even hours of a day). We take as example the case of NDCG@100, but we observed that the relative performance of the ranking methods is equivalent for the other levels of precision. One can notice that, in general, the top three algorithms show a consistent improvement over the simple heuristics *live* and *recency*. The improvement of the *linear log* model over the other two methods is insignificant for `20minutes` – suggesting that the top three ranking solutions are equally effective – but has an important impact for `telegraaf` data set where the improvement is notable for some specific hours (e.g. the improvement for 10 a.m. is 12%.)

3.5.4 An alternative to learning to rank algorithms

A different approach to this ranking problem is to automatically construct a ranking model using learning to rank algorithms. These algorithms propose a straightforward approach to the ranking problem and provide greater adaptability to add more information into the ranking model. We compare the methods based on popularity prediction with several learning to rank algorithms.

Depending on how they address the ranking problem, there are three main classes of learning to rank algorithms: pointwise, pairwise, and listwise [110]. We consider a representative model from each category:

- Multiple additive regression trees (MART) - **pointwise** approach based on the gradient boosting technique proposed in [112].

- RankBoost - **pairwise** approach based on a boosting algorithm and multiple weak rankers [113].

- LambdaMART - **pairwise** and **listwise** approach using boosted regression trees and designed to optimize NDCG [114].

- AdaRank - **listwise** approach also based on a boosting algorithm that minimizes an exponential loss function [115].

Table 3.3: Ranking accuracy in terms of NDCG for different levels of precision. We compare the *linear log* model and the learning to rank algorithms using different set of features: **b**asic and **e**nhanced. The bold value indicates the best performing algorithm for a specific precision level.

(a) 20minutes

Method	NDCG				
	@1	@5	@10	@20	@100
Linear log	**0.35**	**0.52**	**0.56**	**0.58**	**0.61**
MART- b	0.3	0.48	0.52	0.54	0.57
MART - e	0.33	0.50	0.54	0.56	0.59
RankBoost - b	0.07	0.12	0.13	0.14	0.26
RankBoost - e	0.24	0.36	0.39	0.43	0.48
LambdaMART - b	0.06	0.17	0.22	0.25	0.32
LambdaMART - e	0.05	0.16	0.22	0.26	0.32
AdaRank - b	0.13	0.23	0.28	0.31	0.38
AdaRank - e	0.07	0.19	0.24	0.29	0.35

(b) telegraaf

Method	NDCG				
	@1	@5	@10	@20	@100
Linear log	**0.36**	**0.50**	**0.55**	**0.59**	0.60
MART- b	0.31	0.48	0.52	0.56	0.59
MART - e	0.32	0.49	0.54	**0.59**	**0.61**
RankBoost - b	0.19	0.24	0.28	0.32	0.40
RankBoost - e	0.27	0.46	0.49	0.52	0.56
LambdaMART - b	0.13	0.20	0.24	0.30	0.39
LambdaMART - e	0.14	0.21	0.25	0.29	0.40
AdaRank - b	0.15	0.22	0.24	0.24	0.37
AdaRank - e	0.16	0.26	0.41	0.41	0.51

Using the same evaluation strategy (10-fold cross-validation) we deploy and assess the performance of these algorithms for our specific ranking task.[7] While the format of the previous models is not adapted to be used with a large number of features, this can easily be done using learning to rank algorithms. We thus compare the performance of dedicated learning to rank algorithms using the same amount of information as the previous models, with models that include other features into the ranking decision (e.g. section, author, mean inter-comment time). As a result, we train and evaluate these algorithms using two different set of features:

[7]We deploy these algorithms using RankLib open source library [116].

- **basic** set of features: partial popularity, time since publication, publication hour.
- **enhanced** set of features: basic features + (section, author, time of the first comment, mean and median inter-comment time, weekday, and week).[8]

We report the performance of these models in Table 3.3 and compare them with the best performing model from the previous set of tests, the *linear log* model. Overall, one can observe that the *linear log* method is more effective than most of the learning to rank solutions, being surpassed only by the MART model with an enhanced set of features for NDCG@100. From the learning to rank algorithms, MART exhibits effective performances (very close to *linear log* method) across all levels of prediction. This is likely due to the underlying structure of the model that solves the ranking problem through a set of regression trees. Using the basic set of features, the other learning to rank solutions generally do not perform as well as the previous two, which suggest that they are not able to solve the pairwise and listwise constrains for this ranking problem. In general, we observe that adding more features in the model improves the ranking performance except for AdaRank applied to `20minutes` data set, which shows a reduced performance. These results suggest that popularity prediction methods can accurately identify the top most commented articles and could be used as a valuable solution to automatic online news ranking.

3.6 Conclusions

In this chapter, we analyzed the capability to predict the popularity of online news articles. We conducted our study on a large corpus of articles and comments from a French and a Dutch online news platforms and provided insights on how users post comments on news articles. By exploring these data sets we observe that news stories have a very short lifespan and that the volume of comments per article can be described by a power-law distribution. We analyzed the predictive capacity of two content popularity prediction methods and found that a linear model on a logarithmic scale provides the most accurate performance in predicting the popularity of online news articles. In the context online news ranking, we analyzed the ranking effectiveness of two popularity prediction methods and compared them with several baselines methods and learning to rank algorithms. Our results indicate that a linear model on a logarithmic scale is also an effective solution to ranking online news based on their future popularity, with a performance that can evenly match more customized learning to rank algorithms.

[8] Information about *section* and *author* are available only for `20minutes` data set.

Chapter 4

Predictive proactive seeding for mobile opportunistic data offloading

4.1 Introduction

To get real value out of predicting the popularity of web content, we study the effect of this solution in the context of mobile data offloading. In particular we propose the design of a proactive seeding strategy combined with mobile opportunistic communications that can help telecom operators reduce data traffic during periods of increased load.

There are various strategies used by telecom operators to cope with the increasing consumption of mobile data traffic. The typical actions are to optimize the existing network capacity (through better network planning and traffic shaping), to upgrade the network to the next generation technology (e.g., LTE), or to purchase additional blocks of spectrum. More recent alternatives – cheaper and easier to deploy – are built on the notion of mobile data offloading: the use of complementary network technologies to shift in time and space data traffic that is originally intended to traverse the cellular infrastructure.

Mobile opportunistic networks provide a good alternative to offload data with non-real time constraints. By allowing mobile users to access the cache space of collocated users, content requests can be treated through opportunistic communications and thus reduce the data traffic targeted to the cellular infrastructure [12]. Proactive seeding (preloading content into mobile users cache before the actual content request) has often been used in the context of mobile opportunistic communications, where, to reduce the effect of a poor network connectivity, content is preloaded into the cache space of certain mobile users

that can serve as proxies to future content requests [117]. But this strategy can also be valuable for mobile data offloading, where, by anticipating future user requests content can be preloaded in advance during periods of low data traffic to reduce the amount of traffic at future moments of time [118].

The benefit of proactive seeding depends on the capacity to anticipate future user requests. Previously, when similar solutions have been used to reduce the effect of network bottlenecks, predicting the volume of requests was considered a difficult task and simple heuristics have been proposed to detect future popular web objects [89]. Recent findings in the field of social media (as described in Chapter 2) show that the popularity of web content can be predicted and thus improve the impact of proactive seeding. In the following, we elaborate on the results these findings and study the effect of using an actual popularity prediction method as an integrated component of the proactive seeding decission.

4.2 Background

Different strategies have been proposed for mobile data offloading. One practical solution (given the availability and capacity of the resource) is to migrate part of the data traffic from the cellular network to Wi-Fi access points both for the uplink [119] and the downlink traffic [6, 120]. Another approach would be to schedule and preload content into mobile terminals during periods of low data traffic or under better traffic conditions to reduce data traffic in the future. Lee et al. propose a mobile content distribution architecture used to schedule content delivery when the network is lightly loaded and under good physical channel conditions [121]. A technique to reduce the cellular traffic during heavy load has also been proposed by Malandrino et al. [118]. Under the assumption that user-specific requests can perfectly be predicted, the authors propose a water filling algorithm to schedule users' request in advance in such a way that traffic is uniformly distributed across time. The strategy proposed in our work is based on proactive seeding used to reduce the traffic during peak periods. However, in our study we relax the strong assumption of being able to predict user-specific requests and consider that we could only predict the aggregate mobile users content demand. In addition, compared to previous works, we consider that content request delays are tolerated in opportunistic network communications.

The benefit of using opportunistic networks for mobile data offloading has been recently been explored in several recent studies [12, 13, 122–125]. The common scenario considered in these works involve a situation of information flooding, with one message (e.g., web item, file) that needs to be delivered to all other mobile devices. The research question in this case is how to optimize the dissemination of information through opportunistic communications (by making the content available to a larger population of mobile users in the shortest time

duration) and reduce the communication of mobile nodes with the cellular infrastructure. One approach is to use network analysis to detect a set of central nodes (e.g., users that meet many other users), use the infrastructure to push content to these nodes, and rely on the opportunistic communications to further disseminate content to the remaining mobile users [12,122,123]. Another strategy is to control the dissemination progress by reinjecting content into the network when the dissemination though opportunistic communications gets stuck [13,124,126]. We share the same goal – reduce the traffic load though proactive seeding and opportunistic communications – but the situation considered in this work is one where content is heterogenous and the objective is to reduce the data traffic during certain periods of time when the network is heavily used.

Little has been said about the potential use of opportunistic networks for mobile data offloading under the heterogenous data traffic assumptions (i.e., when users show different interest in multiple web items). Li et al. proposed a mathematical formulation of this problem and observed that the greatest offloading potential is obtained when the number of replicas pushed in the network reflect the distribution of content popularity [127]. Similar, Wang et al. proposed an optimal content replication scheme that considers the skewed interest of users in online content [128] and showed that the optimal content replication scheme (to maximize the content requests treated though opportunistic communications) is one that replicates content according to the skewed user' interest. One important limitations of these studies is that they assume predefined and fixed distributions of content popularity and thus ignore the dynamic evolution of content popularity – observed in reality with web content. In our work we assume that the popularity of online content evolves over time and study the impact of different proactive seeding strategies used in the context of mobile data offloading.

4.3 Global scenario

The global scenario considered in this work, and illustrated in Figure 4.1, involves the following three entities: a content producer, a telecom operator, and a population of mobile users.

Content producer: periodically publishes web items, $c \in C$, of fixed size s for a population of mobile users. We consider that the content producer collaborates with the telecom operator with the goal of reducing the communication of mobile users with the cellular infrastructure. The role of the content producer is to track the popularity of each of its web items, with popularity expressed in the number of requests, and to learn models that predict their popularity. Denote $N_c(t_i)$, the number of requests received by a web item c from the time it was published until time t_i and $\widehat{N}_c(t_i, t_r)$ the estimated popularity at time

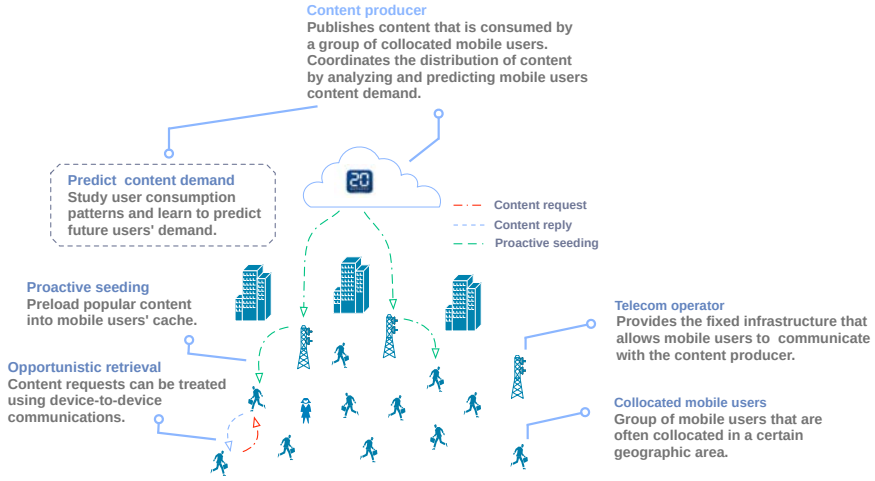


Figure 4.1: The global scenario considered composed of a content producer, located on the Internet, a telecom operator that provides the infrastructure for the communication between mobile users and the content provider, and a group of collocated mobile users.

t_r using the information available up to t_i.

To reduce the number of prediction models (given that time is a continuous variable), we sample time in regular intervals of duration w and learn prediction models only for durations multiple of w. To achieve this, the popularity of online content is recorded during fixed-size time intervals $\{0, w, 2w, ..., t\}$. The smaller value for w the more fine-grained the prediction functions with the downside of an increase in the execution time [129]. In the following we set w to 30 minutes, but the strategy itself makes no assumptions of the sampling rate.

Telecom operator: provides the cellular infrastructure that allows mobile users to communicate with the content producer. In our scenario we consider that the load on the telecom operator comes entirely from the communication of mobile users with the content producer. Concretely, the traffic on the telecom operator side is due to the messages sent from the content producer to the mobile users while content request messages are ignored.

We use the same sampling rate w when analyzing the load on the telecom operator and note L_C^k the number of requests treated by the telecom operator during the interval $[t_k, t_{k+1})$, with $t_{k+1} - t_k = w$. Also, denote L_{max} the maximum accepted load that the telecom operator can handle. We also assume that the telecom operator has a fairly accurate estimation of the traffic load over time, and when $L_C^k < L_{max}$ it can use an additional amount of resources L_{add}^k to proactively seed content to the population of mobile users (under the constraint $L_{add}^k + L_C^k < L_{max}$).

Mobile users: let M be a population of mobile users interested in the web items published by the content producer. To have access to these web items mobile users can either use the cellular infrastructure or by using mobile opportunistic communications (retrieve content from collocated mobile users through device-to-device communications). When using opportunistic communications a content request is broadcasted to the collocated mobile users and, if no answer is received within a delay δ, content is fetched using the infrastructure. We consider that the content producer is notified when a request is treated through opportunistic communications. This is important to maintain accurate statistics about the popularity of each web item and prevents the use of unnecessary resources caused by preloading content already found in mobile users' cache. We also consider that the size of a content c is small enough to be transmitted between two users when they are in direct communication range. The transmission of one content c to a mobile user using the cellular infrastructure takes exactly one unit of cellular traffic (the request itself is negligible). To isolate the effect of a cache replacement policy we also consider that the mobile device storage has an infinite capacity.

4.4 Proactive seeding in mobile opportunistic networks

Proactive seeding operation. We consider that the data traffic that the telecom operator needs to handle per day follows the typical diurnal user activity, with a significant amount of traffic during the day and reduced traffic during the night [130, 131]. The proactive seeding strategy consists in preloading content when the cellular network is less charged ($L_C^k < L_{max}$), by spending an additional amount of traffic (L_{add}), with the goal of reducing the network load at a further moment in time. To achieve this we consider a cooperation between the telecom operator and the content producer where the decision of when to preload content is taken by the telecom operator based on its traffic load statistics and the decision of what to preload (which web item and how many replicas) is the decision of the content producer.

4.4.1 Premise for effective proactive seeding

In our scenario, the value of proactive seeding is measured in the volume of traffic that is reduced from the traffic peak periods. Intuitively, the more content is preloaded during idle periods the greater the reduction of data traffic at later moments. But there is an important danger in proactive seeding: if the speculations made on the future user actions are incorrect the outcome can be an inefficient – or even an extra – utilization of network resources. From a more general point of view we distinguish the following dimensions that can influence the value of proactive seeding in a mobile opportunistic network scenario:

- *Network resources*: the goal of proactive seeding is to level out the data traffic by using the network resources during periods of low user activity. An inefficient use of these resources can lead to an additional load (and extra costs) on the telecom operator facilities.

- *Mobile device resources*: preloading content into mobile terminals can also consume additional resources at the mobile terminal side (e.g., battery and processing power, storage capacity, and quota from the data plan).

- *Content characteristics*: the properties of online content (distribution of popularity, lifetime, or size) can also affect the value of proactive seeding. For example by knowing that distribution of popularity is highly skewed (e.g., Zipf-type distribution) one can tune the seeding decision to identify and promote only the most popular web items.

- *Quality of the predictions*: the impact of proactive seeding depends on the capacity to correctly predict future users' requests. Accurate predictions will translate in a decrease of traffic during periods of traffic peak; wrong guesses will lead to inefficient use of handsets and network resources. Prediction thus plays a crucial role in this process and can incline the balance from reduction of data traffic to inefficient use of resources.

- *Opportunistic mobile contacts*: data about user mobility can be capitalized into useful information in the proactive seeding action. For example, by knowing that two users, interested in the same content, will be in direct communication range, content can be preloaded to one of the users that could further transmit it to its peer.

Proactive seeding is thus a complex process and, while a study of the inter-play of these aspects can be very useful, in this work we focus on one particular aspect of the problem: the quality of the predictions. In particular we compare how the different levels of predictions can reduce the data traffic during periods of increased load.

4.4.2 Proactive seeding strategies

We consider the following strategies for proactive seeding:

- *Constant* popularity: this method considers that the popularity of online content remains stable over time and thus the proactive seeding decisions based on the popularity of a web item at time t_i, $N_c(t_i)$. This approach has been used to as a solution to prefetch web pages [89] but also in the context of proactive seeding in opportunistic networks [117].

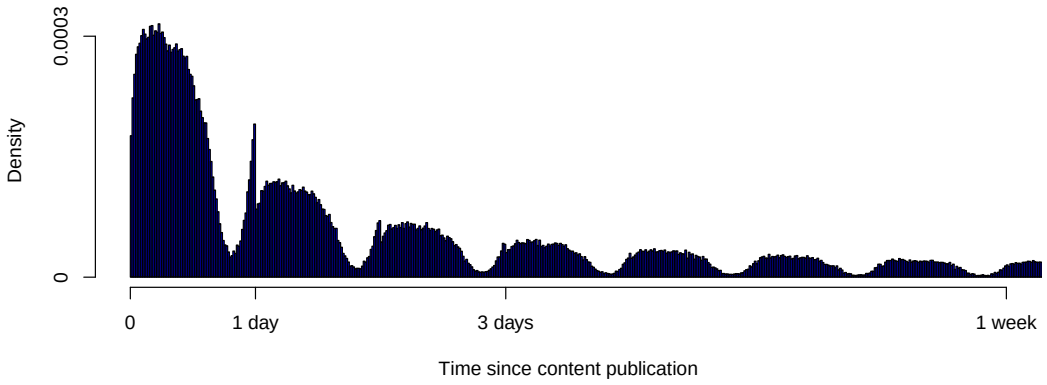


Figure 4.2: The probability distribution function for the request arrival times relative to the content publication time for MediSyn synthetic workload. We represent the histogram covering a one-week period.

- Predicted popularity (*linear log* model): under this strategy we predict the actual number of requests that a content will receive at a future time t_r ($t_r > t_i$), $\widehat{N}_c(t_i, t_r)$. To estimate this value we use a linear model on a logarithmic scale, that, as we presented in Chapter 3, showed the most accurate performance in predicting the popularity of online news.

- Perfect popularity prediction (*perfect prediction*): because predicting the popularity of online news articles is prone to errors, we also consider the case where the popularity prediction algorithm can perfectly predict the popularity of web content, $N_c(t_i, t_r)$.

- Proactive seeding *scheduler*: this method assumes a perfect knowledge of user-specific requests (it knows exactly when web items will be requested by mobile users) and schedules the seeding decisions to level out the traffic throughout the day. This strategy has been considered as a fine-grained solution in proactive seeding [118], and although nowadays such a granular level of prediction would be difficult to attain, this could be achieved in the near future through a better understanding of the social influence and the information diffusion in social networks [21–23].

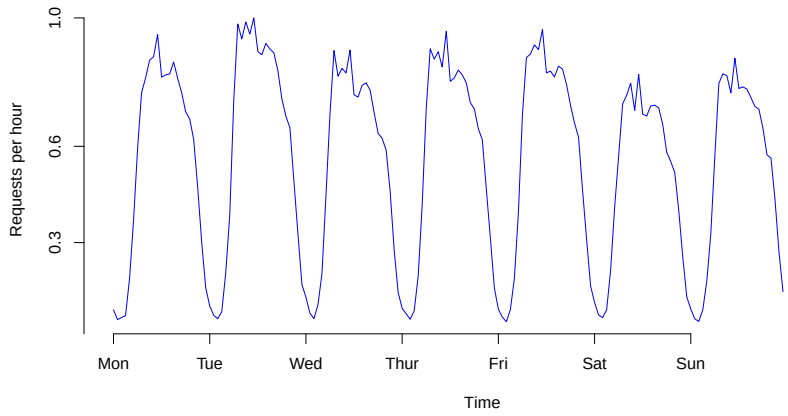


Figure 4.3: The distribution of content request per hour, on a weekly basis, generated by MediSyn.

4.5 Evaluation

4.5.1 Simulating user behavior

To study the benefit of proactive seeding, and in the absence of a data trace that contains both mobility and content request information, we combine traces that depict user mobility in real-life scenarios with traces that simulate HTTP requests. To simulate user mobility, we rely on two real-life connectivity traces and use a synthetic HTTP request workload generator to simulate users' content requests.

Simulating content requests. To simulate mobile content request patterns we use Medisyn, a synthetic workload generator [132]. Most of the existing workload generators consider a predetermined and fixed popularity over time, which, as we saw in Chapter 3, is unrealistic as online content has a limited lifetime and its popularity decreases over time. MediSyn is a workload generator that reproduces more closely the properties observed with real HTTP traffic as it is designed to simulate the dynamic evolution of content requests over time. To create more realistic content request patterns we tune the parameters of the workload using the empirical observations made on news articles described in Chapter 3. We summarize the main parameters of the simulation in Table 3.1.

Using these configuration parameters we generate workload for a duration of one month. We illustrate two properties for the resulting workload trace: the non-stationarity of content popularity (Figure 4.2) and diurnal content request patterns (Figure 4.3). In Figure 4.2 we

Table 4.1: Summary of the content creation and users' requests.

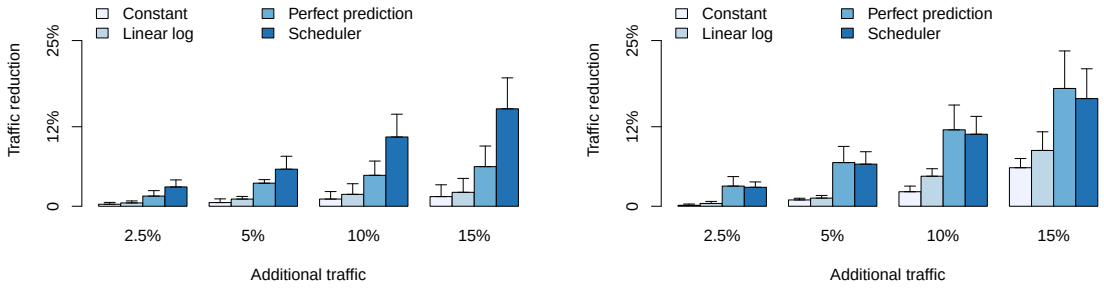
Component	Model
Content creation:	
- New contents per day:	Pareto ($\alpha = 1.13$)
- Content arrival process:	Pareto ($\alpha = 1.01$)
Content request:	
- Popularity:	Zipf ($\alpha = 1.2$)
- Content life span:	Log-normal ($\mu = 5.74$, $\sigma = 1.22$)
- Request arrival process:	Non-homogenous Poisson process

represent the probability density function of content request times relative to content publication time. We observe that the most significant share of requests (on average 60%) are received in the first day after the publication and the probability to receive requests drops considerably after one week. In Figure 4.3 we represent the average number of content requests during one week and per hourly basis. Even if the workload generator does not perfectly reproduces user activity on an hourly basis (compared to what has been observed with online news articles in Figure 3.1) the circadian patterns of users requests are nevertheless fairly approximated: users' activity starts to increase around 7 a.m., it presents the most intense activity between 11 a.m. and 5 p.m., and it reveals a reduced activity during the night.

Table 4.2: Mobility datasets characteristics.

Dataset	*Number of participants*	*Trace duration*	*Probing interval*	*Type of activity*
Rollernet	61	1h30	15s	Sport
Stanford	200	1h	20s	Scholar

User mobility. To simulate user mobility we use two real-life contact traces, *Rollernet* and *Stanford*. *Rollernet* is a contact mobility trace that describes the connectivity characteristics of 62 participants captured during a rollerblading tour in Paris for a duration of one hour and a half [133]. We also consider *Stanford* mobility trace in our evaluation, a data set that reflects user mobility in a different setting [134]. This trace allows us to replay the contacts between 788 individuals (students, teachers, and other members) during a typical school day (between 7 a.m. and 5 p.m.) at an American high-school. This trace describes users' daytime mobility for a longer duration of time and on a larger population of users. We summarize the characteristics of the two traces in Table 4.2.



(a) *Stanford* (b) *Rollernet*

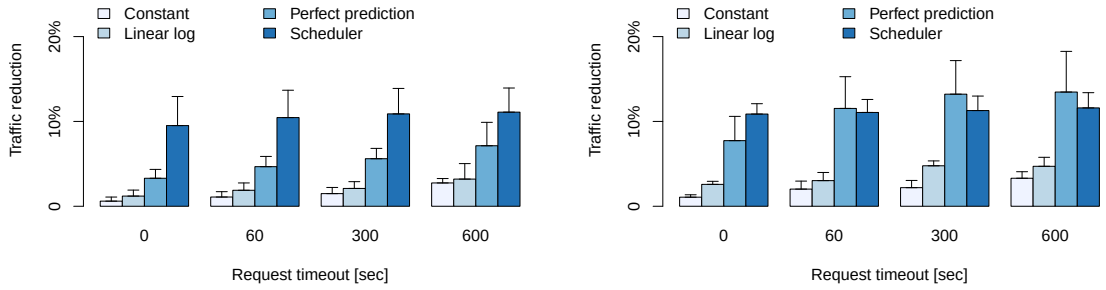
Figure 4.4: The performance of the different proactive seeding strategies using a request timeout of 60 seconds and for different values of the additional load.

4.5.2 Simulation scenario

To jointly simulate users' content requests in a mobile scenario we map the HTTP request trace to the users of a mobility trace such that each content request is randomly assigned to a mobile user. Users consume content at any moment of time (according the workload characteristics) but they are collocated only during a limited period of time (according to the characteristics of a mobility trace) and separated otherwise. Taking the example of *Stanford* data set, users appear in the mobility setting around 7 a.m. and leave the mobility setting after 5 p.m.. Given that the scenario described by the *Rollernet* data set covers a much reduced period of time we replay the trace several times to cover the same duration as *Stanford*.

For the HTTP request trace used in this work, traffic is significantly higher during the typical working hours (from 9 a.m to 7 p.m.). We thus measure the effectiveness of proactive seeding in the percent of traffic that is reduced from the periods of increased data traffic (9 a.m. – 7 p.m.). For the popularity-based methods the additional credit, L_{add}, is proportionally split according to the estimated future demand [1]. To replicate the content, we assume that the content producer has no indication about users interest in web content, and therefore it randomly select mobile users in the proactive seeding process. We consider an additional average traffic of 2.5%, 5%, 10%, and 15% (average, because depending on the workload this value may vary) and consider different values for the request time-out period δ: 0, 60, 300, 600 seconds.

[1] We have also tried a square-root replication scheme, that is known to be the optimal allocation scheme in unstructured Peer-to-Peer networks [135], but the results were less efficient.



(a) *Stanford* (b) *Rollernet*

Figure 4.5: The performance of the different proactive seeding strategies for an additional traffic of 10% and for different durations of the request timeout.

4.5.3 Results

In Figure 4.4 we illustrate the traffic reduction when applying the different proactive seeding strategies. The results show that, when considering that the popularity of online content remains constant over time the benefit of proactive seeding is poor. For example, for an additional traffic of 10% used during idle periods the traffic reduction during the day is, on average, 1% for *Stanford* and 2% for *Rollernet*. On the other hand, the ability to predict future content demand and adjust the proactive seeding decisions accordingly, has a greater potential in traffic reduction. For instance, by learning an actual popularity prediction model on the history of user requests one can double the amount of traffic that can be reduced compared to the model that considers that the popularity remains constant over time. Moreover, the benefit that can be obtained using a prediction algorithm that perfectly knows the future shows a 5 times potential increase. Thus, even if for this type of workload the benefit of using an actual prediction model appears to be limited, the theoretical improvement that can be obtained with a more accurate prediction method (or using a data set with better predictive characteristics) is a good indicator of the benefit brought by content popularity prediction methods.

The previous set of results considered a request timeout of 60 seconds, which, in a disconnected mobile setting may limit the impact of proactive seeding. To observe how the performance is influenced by the duration of the request timeout, we vary this value from 0 to 600 seconds while keeping the additional traffic to10%. The results show that even for a request timeout of 600 seconds the potential reduction obtained for the *constant* popularity solution is of only 2.7% for *Stanford* and 3.3% for *Rollernet*, whereas, when using an actual prediction algorithm the performance reaches 3.7% for *Stanford* and 4.7% for *Rollernet*. Thus, under the *constant* popularity assumption, by promoting stale content into the network, this solution shows that it is unadapted for the practical use because it

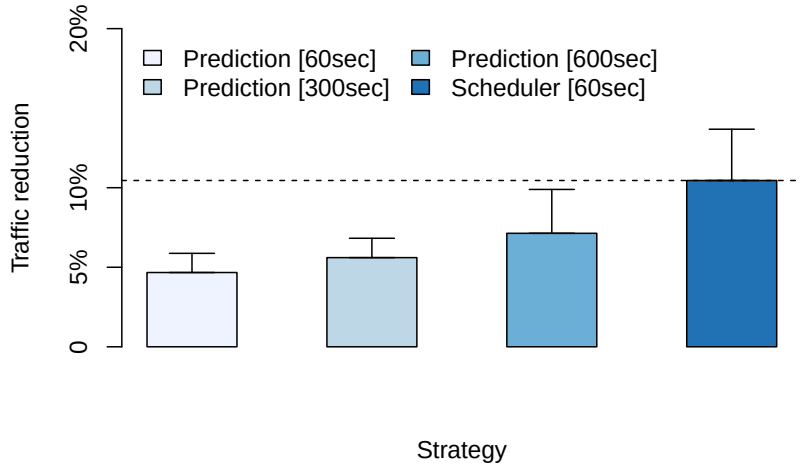


Figure 4.6: For an additional credit of credit of 5%, the performance of the *Scheduler* strategy using a request timeout of 60 seconds, compared to the perfect prediction strategies under different values of request timeout.

wastes important network resources with little gain in terms of the future traffic reduction.

Certainly, relying on a solution that predicts individual user requests (in our case the *Scheduler* method) guarantees an efficient use of the network bandwidth if the predictions are perfect. However, this altruistic manner of preloading content may not be the most effective solution. This can be observed for the *Rollernet* scenario, where, by planning the proactive seeding decisions based on the global popularity demand one can obtain an average improvement of 3% compared to the *Scheduler* strategy. *Stanford*, on the other hand, shows a different behavior with the *Scheduler* strategy showing an improvement of 5% over perfect prediction method. The different performance observed in the two cases is explained by the connectivity characteristics of the mobile traces: *Rollernet* represents a more dynamic and connected mobile environment compared to *Stanford* and thus requests are better treated through opportunistic contacts. In poorly connected mobile environments the benefit of a popularity-based proactive seeding can, nevertheless, be improved by increasing the duration of the request timeout. Taking the example of *Scheduler* method for an additional 5% traffic and a request timeout of 60 seconds, by increasing the duration of the request timeout, one can reduce the performance gap between the two strategies (Figure 4.6).

4.6 Conclusion

In this chapter we proposed the use of content popularity prediction methods to improve the efficiency of proactive seeding in the context of mobile opportunistic data offloading.

Compared to traditional strategies that consider a stable evolution of content popularity over time, the strategy used in this case is to actually predict future content demand and adjust the proactive seeding decisions accordingly.

To evaluate the benefit of this solution in a real-life deployment, we proposed a simulation scenario that reproduces, to a certain extent, the mobility and content request characteristics of a group of collocated mobile users. In this scenario the objective is to reduce the amount of traffic that the mobile users create during the day by preloading content when the network is less loaded.

A preliminary set of results show that proactive seeding can have a greater impact if the decision of what content to replicate is based on an algorithm that predicts future content demand. Although, for this particular workload trace, the benefit of using an actual prediction algorithm is limited, the theoretical gain obtained under the assumption that the global popularity can better be predicted is a strong evidence of the potential value of using this strategy.

Chapter 5

Beyond contact predictions in mobile opportunistic networks

5.1 Introduction

The design of efficient communication protocols in mobile opportunistic networks depends in great part on the capacity to understand human mobility characteristics. Over the last years several studies have revealed important insights about the duration of contacts and inter-contact between mobile users [8, 136, 137], the periodicity of these encounters [9], or the network structures created by human interactions [10, 138–140]. In the context of mobile opportunistic networks, uncovering mobility patterns can then be used to design measures that facilitate the prediction of contacts between mobile users. This includes the use of frequency of contacts to identify similarities between mobility characteristics [141], or in finding strongly-connected mobile users that could serve as message carriers [142]. While these metrics can serve as good heuristics to predict contacts between mobile users, they have a limited power in detecting future contact opportunities. A more advantageous but laborious approach to this problem is to actually train a model that can predict future contacts between mobile users.

Recent studies have addressed the problem of contact prediction – predict if two nodes are going to be in direct transmission range – and have revealed that, under the right prediction method and predictive features, contacts between mobile users are, to a certain extent, predictable [143]. This result is valuable as it allows one to actually predict human encounters and design more effective communication protocols.

The properties and the impact of κ-vicinity view in mobile opportunistic networks has been studied by Phe-Neau [144].

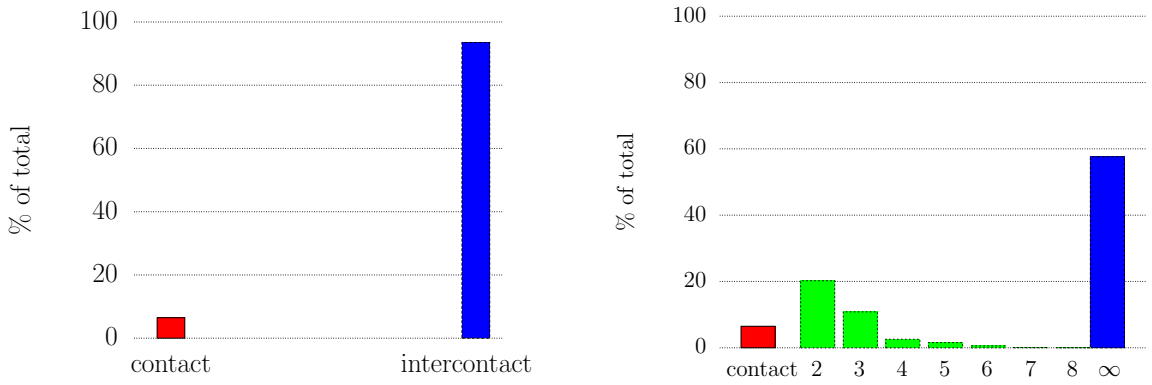
But the pairwise relationships between mobile users can be described by more than the binary (contact / intercontact) view as often, individuals may find themselves not in direct transmission range but in the nearby vicinity. Thus, to have a more comprehensive view on the available communication opportunities, the extended notion of contact, namely κ-contact, has recently been proposed [145]. Previous analyses showed that considering only contacts between mobile users offers a biased and suboptimal network understanding while studying κ-contacts provides a more complete understanding of the available end-to-end communication opportunities.

In this chapter, we provide novel insights about the κ-contact relationships and show that considering only direct contacts provides us a limited view about the pairwise communication possibilities. We then study the predictability of κ-contacts. Using data from three human-based contact traces, we compare the accuracy of predicting κ-contacts with the traditional case of predicting direct contacts between mobile users. We show that κ-contact opportunities are more predictable than direct contact relationships. To measure the possible impact of these findings in a real-life application we analyze the impact of using a κ-contact prediction model as a solution for mobile data offloading. Through simulation we show that there is a greater potential of relying on κ-contact prediction compared to the traditional contact case.

5.2 Background

Understanding human interactions and mobility has been the main subject of several recent studies. Song et al. observed that, despite many decisions influencing mobile users' daily routines, there is a high degree of predictability in user mobility (an average 93% potential) with low variability across the population [146]. Clauset and Eagle revealed strong periodicities in contact periods between mobile users which depend on the environment under study (the physical place and the type of user activity) [9]. Zayani et al. studied the problem of predicting contact opportunities between mobile users [143]. Using a tensor-based link prediction method, the authors analyze the predictive power of various features that capture both the topological distance and the physical proximity between users. In this work we focus on one specific aspect of human mobility, i.e., predicting if two nodes will be in each other κ-vicinity. Our analysis is close to the work of Zayani et al. but differs in the prediction goal (we extend the contact prediction to the κ-contact case), the connectivity traces under study, and the prediction framework (we use a supervised learning framework compared to the unsupervised setting used in their work).

From a general point of view, the prediction objective presented in this work is related to the link prediction problem studied in the context of complex networks. This topic



(a) Current vision.

(b) Vicinity-aware vision.

Figure 5.1: *Sig09* example: current vision versus vicinity awareness.

has been an active research direction in several domains that define relationships between different entities. This includes predicting the co-authorship of research publications, hyperlinks between web pages, or human communication activities [147–150]. Liben-Nowell and Kleinberg studied the predictive power of various topological features and observed that the Katz measure performs consistently well [147]. While analyzing the predictive power of non-topological features Al Hasan et al. observed that the frequency of interactions (e.g., the number of times two people co-authored scientific papers) is an efficient predictive variable for future interactions [151]. We build on these findings, and rely on the predictive power of various features (topological measures and the frequency of mobile users' encounters) to predict the κ-contact opportunities.

5.3 Vicinity and data sets

5.3.1 Beyond contact relationships

Previous studies on mobile opportunistic networks considered only network knowledge coming from nodes in contact. This approach has proved to be a good-enough solution in making forwarding but it has its limitations in what concerns the network view about the end-to-end communication opportunities. Let us take for example two mobile users tracked during Sigcomm 2009 conference (entitled *Sig09* in the following) and look at the proportion of time spend by nodes at a certain distance from one another. Using the binary view, we observe in Figure 5.1(a) that the two mobile users remain 6% of the time in contact and the remaining 94% in intercontact; thus we are inclined to say that there is a weak communication potential between this pair of nodes. However, when we analyze the same situation from a vicinity-aware point of view (see Figure 5.1(b)) we observe that the two

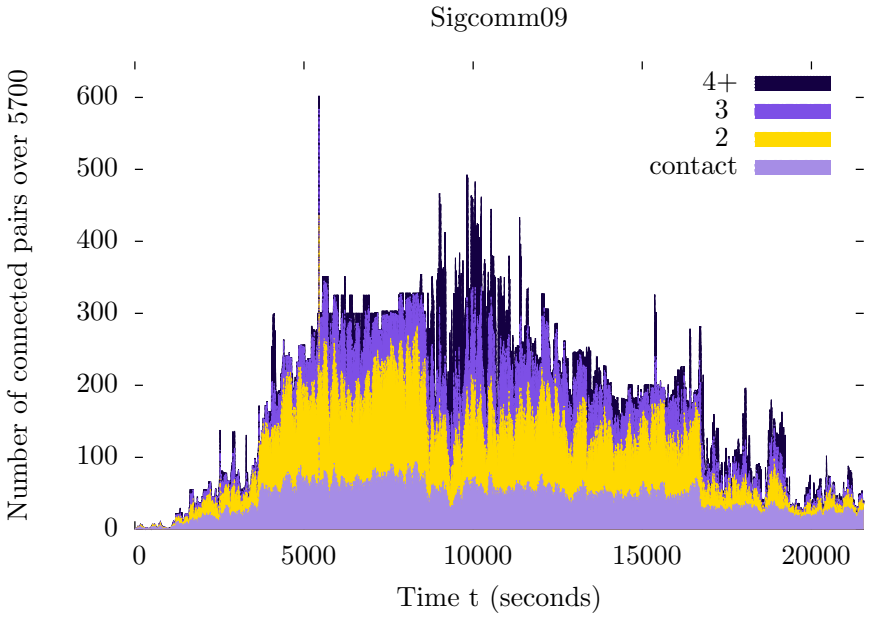


Figure 5.2: *Sig09* end-to-end transmission opportunities.

nodes also remain at a 2-hop distance around 20% of the time and at a 3-hop distance around 10% of the time. Seen from this perspective the proportion of time nodes spend without any end-to-end path linking them (∞) is 57% of the duration; far below the 94% intercontact duration illustrated in Figure 5.1(a).

In addition, from a general point of view we observe that direct contacts represent only a fraction of the available end-to-end transmission opportunities. Taking again the example of *Sig09*data set, in Figure 5.2 we represent the number of pairs connected by a their shortest distance. The bottom layer indicates the number of pairs in contact, the yellow layer shows nodes connected by 2-hop paths and so on. We can see from this example that most end-to-end transmission opportunities come from 2-hop paths and not from direct contacts. Thus, viewing the pairwise relationships beyond the contact / intercontact paradigm improves our understanding about the available end-to-end communication opportunities.

5.3.2 κ-vicinity, κ-contact, and κ-intercontact

To characterize the notion of vicinity in DTN, we adopt the the concept of κ-vicinity proposed by Pheneau et al. [152]. We discriminate a node i's vicinity according to the number of hops between i and its surrounding neighbors. We also assume that connectivity is bidirectional which makes κ-vicinity relationships symmetric.

Definition 1 ***κ-vicinity.*** *The κ-vicinity $\mathcal{V}_\kappa^i$ of node i is the set of nodes with shortest*

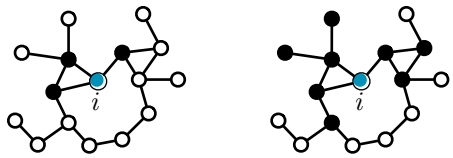


Figure 5.3: Example of κ-vicinity. The 1-vicinity consists in in all nodes found at a 1-hop distance. The 2-vicinity consists in all i neighbor's whose shortest distance is less than 2 hops.

paths of length at most κ hops from i.

As an example, we illustrate in Figure 5.3 the 1-vicinity and 2-vicinity for node i. Clearly, $\mathcal{V}^i_{\kappa-1} \subset \mathcal{V}^i_\kappa$.

Definition 2 ***κ-contact.*** *Two nodes are in κ-contact when they dwell within each other's κ-vicinity, with $\kappa \in \mathbb{N}^*$. More formally, two nodes i and j are in κ-contact when $\{i \in \mathcal{V}^j_\kappa\} \equiv \{j \in \mathcal{V}^i_\kappa\}$. In other words, a contemporaneous path of length at most κ hops links i and j. Note that, 1-contact represents direct contact.*

Definition 3 ***κ-intercontact.*** *Two nodes are in κ-intercontact when they do not belong to each other's κ-vicinity (there is no path of length κ or less linking the two nodes).*

5.3.3 Data sets

Table 5.1: Data sets characteristics.

Data set	#	*Duration*	*Probing*	*Type*
Infocom05	41	12h	120s	Conference
Sig09	76	1 day	120s	Conference
Rollernet	61	1h30	15s	Sport

We consider several real-life contact traces throughout our experiments.

Infocom05 measurement was held during a 5 days conference in 2005 [8]. 41 attendees carried iMotes collecting information about other iMotes within a 10m wireless range. We study a 12-hour interval bearing the highest networking activity. Each iMote probes its environment every 120 seconds. *Infocom05* represents a professional meeting framework.

Table 5.2: The average duration that nodes remain at a certain distance (in seconds).

	κ						
Data set	1	2	3	4	5	6	7
Infocom05	399	296	224	175	131	154	212
Sig09	149	83	41	25	18	13	11
Rollernet	48	65	76	89	105	114	129

Sig09 trace was captured during the first day of Sigcomm 2009 conference in Barcelona [153]. The experiment tracked 76 users using Bluetooth-based smartphones. Each phone probed the environment every 120 seconds to log all users in direct communication range.

Rollernet had 62 participants measuring their mutual connectivity with iMotes during a one hour and a half rollerblading tour in Paris [133]. For this experiment the iMotes were configured to scan the environment every 15 seconds. This experiment shows a specific sport gathering scenario.

In Table 5.1, we recapitulate all data sets characteristics: # is the number of participating nodes; *Duration* indicates the data set duration; *Probing* shows the probing intervals of the measuring devices.

5.4 Pairwise relationships under the κ-contact case

Given the new definitions of κ-contact and κ-intercontact we analyze different characteristics of the pairwise interactions. (For additional information concerning κ-contact and κ-intercontact properties, please refer to [152].)

5.4.1 Pairwise minimum distance

We begin by studying the pairwise minimum distance, i.e., how close nodes come to each other throughout the duration of a trace. For instance, if two nodes meet at least once, we mark this distance as 1. If they come as close as 3 hops, we consider the minimum distance to be 3. For nodes that never come in κ-contact, we consider this distance as ∞.

We represent the results in Figure 5.4. In terms of pairs of nodes that come in direct contact, we observe that in conference settings, characterized by a high number of nodes in restricted physical spaces, the number of connected pairs is reasonably high: 49% for *Sig09* and 73% for *Infocom05*. *Rollernet* on the other hand shows a lower network connectivity, with only 33% of nodes coming in a direct contact. But the analysis of contact alone yields an incomplete picture as there is a considerable amount of nodes who come close to each other but never in direct communication range. For example, the percentage of pairs that come at a distance of 2 is 5% for *Infocom05*, 16% for *Sig09*, and 41% for *Rollernet*.

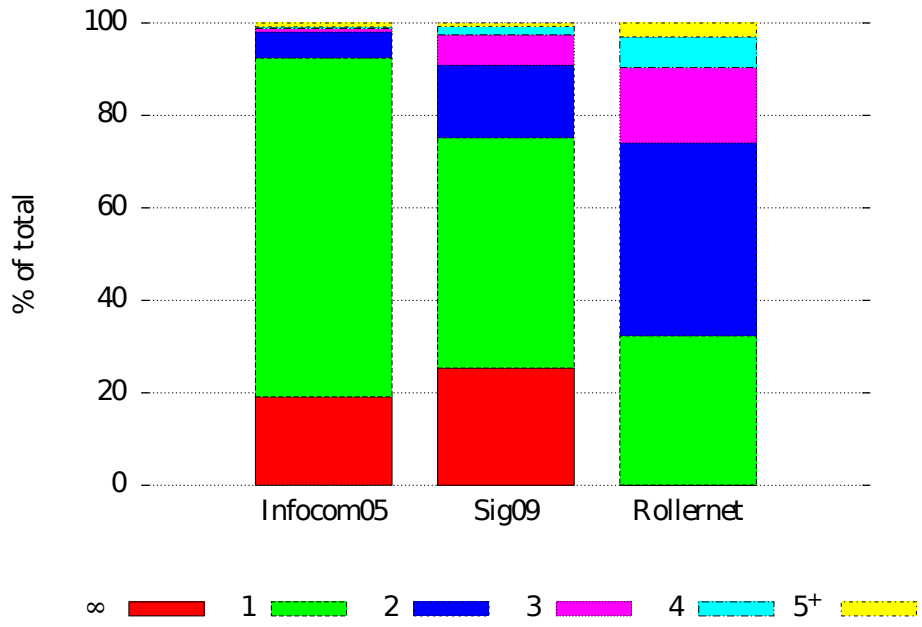


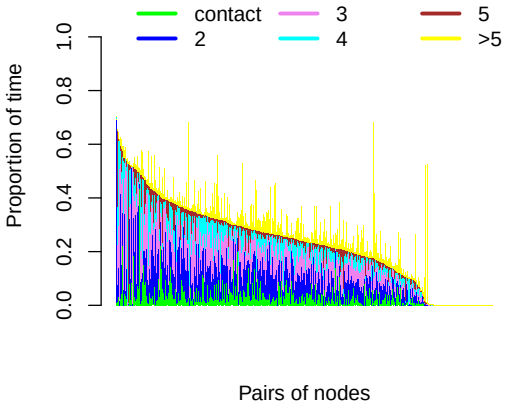
Figure 5.4: Pairwise minimum distance for *Infocom05*, *Sig09*, and *Rollernet*.

For *Rollernet* the percent of nodes that come at a 2-hops distance is even higher than the nodes that come in direct contact and one can observe that a non negligible amount of nodes advance up to a distance 3 (16%) and 4 (6%).

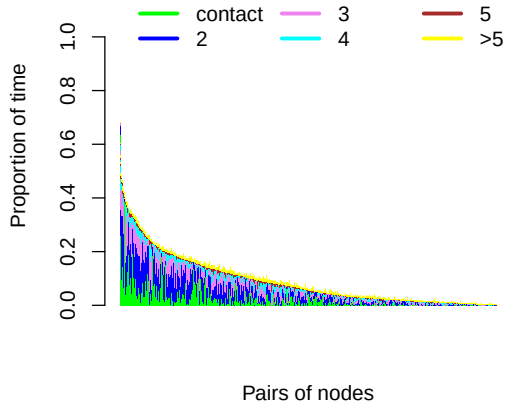
5.4.2 Analyzing the distribution of pairwise distance

In a mobile scenario the distance between nodes changes over time due to nodes' movement. To understand the repartition of time that nodes spend at a certain distance from one another we analyze the distribution of pairwise distance. The results are presented in Figure 5.5, with pairs of nodes ordered by how long they stay in κ-contact $= 5$ and where we delineate the proportion of time spend at each distance.

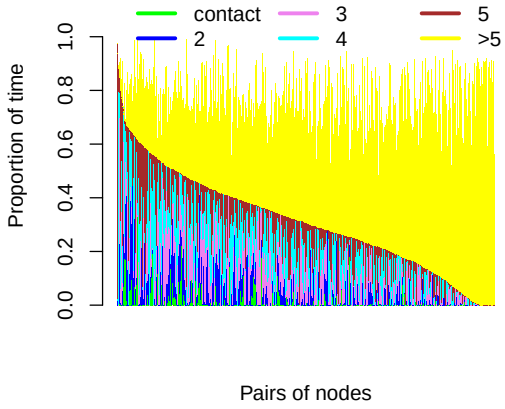
The three data sets depict different connectivity characteristics. *Sig09* shows a poor network connectivity with, on average, 93% of time nodes finding themselves in intercontact. *Rollernet*, a denser and more dynamic mobility setting, reveals stronger connectivity characteristics with nodes finding themselves in intercontact only 37% of time. Figure 5.5 also illustrates that direct contacts are a scarce resource. On average, users spend only 2.1% (*Rollernet*-1.5%, *Infocom05*-3.2%, *Sig09*-1.6%) of their time in contact and a more significant amount of time in the close neighborhood, e.g., 4.7 % at a distance 2 (*Rollernet*-4.9%, *Infocom05*-6.9%, *Sig09*-2.4%), 5.3 % at a distance 3 (*Rollernet*-7.5 %, *Infocom05*-6.8 %, *Sig09*-1.6%), and 4.4 % at a distance 4 (*Rollernet*-8.8 %, *Infocom05*-4.2 %, *Sig09*-0.7%).



(a) *Infocom05*



(b) *Sig09*



(c) *Rollernet*

Figure 5.5: The proportion of time (relative to the duration of a mobility trace) that nodes spend at a certain distance from one another. To simplify the representation we delineate with a distinct color only the first five distances and represent the remaining distances under a unique color (yellow). The pairs are ordered by the proportion of time spent in κ-vicinity = 5.

But even users that come in contact spend only a limited proportion of time in direct communication range. On average, nodes that meet at least once throughout the duration of a trace, spend only 3.9% of time in contact (*Sig09*-3.1%, *Infocom05*-4.1%, *Rollernet*-4.5%) and a non negligible proportion of time in their immediate vicinity: 7.2% at a distance 2 (*Sig09*-4.6%, *Infocom05*-9%, *Rollernet*-8%), 6.5% at a distance 3 (*Sig09*-2.7%, *Infocom05*-8%, *Rollernet*-9%), and 5% at a distance 4 (*Sig09*-1.1%, *Infocom05*-5%, *Rollernet*-9%).

Table 5.3: The average duration of a κ-contact relationship (in seconds).

	κ						
Data set	1	2	3	4	5	6	7
Infocom05	399	322	274	247	230	224	224
Sig09	149	101	72	60	54	51	50
Rollernet	48	61	68	75	81	86	90

Thus even if users have the tendency to get further from one another, they often remain in the immediate vicinity.

5.4.3 The stability of κ-contact relationships

The previous analysis gives us a better understanding of the proportion of time spent by nodes at a certain distance from one another but it provides limited information about the frequency of change of these distances over time. We thus analyze how often the distance between two nodes changes and how often nodes leave each other κ-vicinity.

In Table 5.2, we present the average duration of an interval during which nodes remain at a distance of κ-hops from one another. For *Infocom05* and *Sig09*, we observe that close connections are more stable, with smaller average durations as the distance between nodes increases. This shows that for conference settings, network stability comes from the core of the κ-vicinity. However, we observe the opposite phenomenon for *Rollernet* data set. With larger κ we have an increase of the average duration that nodes spend at a certain distance from one another. Thus, due to nodes' movement in a highly dynamic scenario, meeting between users lasts for very short periods of time but nodes spend a significant amount of time in the nearby vicinity.

We also study the average κ-contact durations (see Table 5.3), i.e., we observe the average duration of each κ-contact interval. Intuitively we would expect that, since we cover a wider spatial range with our κ-vicinity, nodes coming closer are likely to be in κ-contact earlier and leave the κ-contact later, therefore we should obtain longer κ-contact intervals. With *Rollernet*, we observe that the greater the value for κ, the longer the durations. Surprisingly for *Infocom05* and *Sig09*, this is not the case, we actually notice the opposite phenomenon. With larger κ, we seem to have smaller κ-contact intervals. So does that mean that increasing our network vision with the κ-vicinity reduces the duration of end-to-end transmission possibilities?

Table 5.4 shows how wrong this conclusion may be. In this table, we show the actual number of κ-contact intervals for each κ and each data set. For all of them, the greater the value of κ, the greater the number of κ-contact intervals. So, with higher κ values, we multiply the possibility of observing a κ-contact interval. They may be on average of

Table 5.4: κ-contact number of intervals ($\times 1,000$).

	κ						
Data sets	1	2	3	4	5	6	7
Infocom05	3.7	14.7	28.9	40.0	46.7	50.3	51.9
Sig09	13.3	49.7	96.9	131.6	152.2	163.4	168.8
Rollernet	2.6	9.4	18.4	27.5	35.2	41.3	45.7

shorter length (for *Infocom05* and *Sig09*) yet we multiply the possibility of having pairwise end-to-end paths. In addition, the cumulated κ-contact duration grows with larger κ. A similar observation as well as an explanation has been made in a companion paper [152].

5.5 Predicting κ-contact encounters

5.5.1 Dynamic graph representation

The mobile traces analyzed in this paper represent dynamic networks composed of a set of mobile users that sporadically come in contact. We represent this network using a dynamic graph structure, $G_{0,T} = (V, E_{0,T})$, with V the set of mobile users observed during a finite period of time [0, T) and $E_{0,T}$ the set of temporal edges between them. We consider an edge $e_{uv} \in E_{0,T}$ if any two users $u, v \in V$ have been at least once into contact during the period [0, T). To analyze the evolution of this network over time, we split time into fixed time-windows of duration w and represent the dynamic network as a time series of network snapshots G_{t_1}, G_{t_2}, ..., G_{t_n}, with $n = \lceil \frac{T}{w} \rceil$. G_{t_i} represents the aggregate graph G_{t_{i-1},t_i} that records the contacts between mobile users during the period $[t_{i-1}, t_i)$. In a dynamic network, the future changes of the network may depend not only on the most recent state of the network but also on older ones. To model the dynamic evolution and catch possible periodicities in human encounters, the data used as input in the prediction process is represented as a successive series of static snapshots $G_{t_{i-m}}, ..., G_{t_{i-2}}, G_{t_{i-1}}$. Thus, given data from the previous m time-windows our objective is to predict the κ-contacts during the next target period G_{t_i}. We will later discuss how the choice of w and m affect the prediction performance.

5.5.2 κ-contact prediction problem

We formulate the prediction task as a binary classification problem where, given past data recorded until a moment in time t_{i-1}, the goal is to predict if any two mobile nodes will be in κ-contact during the subsequent period $[t_{i-1}, t_i)$.

We rely on two types of information in the prediction model: the frequency of κ-contact occurrences and the structural properties of the connectivity network. The first

type of information measures the strength of κ-contact relationships, expressed by the *duration* and the *number of times* any pair of nodes has been in κ-contact in the past. A longer duration and a greater number of κ-contacts can provide stronger evidence that two nodes will be in κ-contact in the future. For the second type of information, to quantify the structural properties of the network, we extract various features that capture the proximity between nodes in the network of past interactions. These features showed strong predictive power in various prediction tasks such as collaborative filtering and link prediction problems [147, 148, 154]. In this work we use the following proximity measures:

- *Common neighbors (CN).* For each pair of nodes $u, v \in V$, CN represents the number of common neighbors:

$$CN_{(u,v)} = \mid \mathcal{V}_1^u \cap \mathcal{V}_1^v \mid. \tag{5.1}$$

- *Adamic Adar* [155]. This measure extends the notion of common neighbors by weighting each neighbor by the inverse logarithm of its degree centrality:

$$AdamicAdar_{(u,v)} = \sum_{x \in \{\mathcal{V}_1^u \cap \mathcal{V}_1^v\}} \frac{1}{\mid \mathcal{V}_1^x \mid}. \tag{5.2}$$

- *Katz* [156]. This feature counts all the paths between any pair of nodes, giving a higher weight to shorter paths. If $path_{u,v}^l$ represents the set of paths of length l between two nodes u and v, and β is a damping factor (set to 0.05 in our evaluation), the Katz score is calculated using the following formula:

$$Katz_{(u,v)} = \sum_{l=1}^{\infty} \beta^l \times \mid path_{u,v}^l \mid. \tag{5.3}$$

- *Preferential attachment* [157]. This feature is built on the premise that the probability of a new contact is correlated with the product of nodes' degree.

$$PA_{(u,v)} = \mid \mathcal{V}_1^u \mid \times \mid \mathcal{V}_1^v \mid \tag{5.4}$$

The two types of features provide complementary information about pairwise κ-contact characteristics. The frequency of interactions catches the persistence of κ-contact relationships but its predictive power is conditioned by the past contact occurrences (using these features one can only predict the reoccurrence of a pairwise κ-contact relationship). Topological features, on the other hand, allow us to capture complex data patterns about the structure of the network of interactions. We build the prediction model and report the

Table 5.5: Notation for the binary classification confusion matrix

		Predicted value	
		predicted = 1	predicted = 0
Actual value	actual = 1	TP	FN
	actual = 0	FP	TN

results using the entire set of features as we observed that taking these features together achieves the most accurate performance.

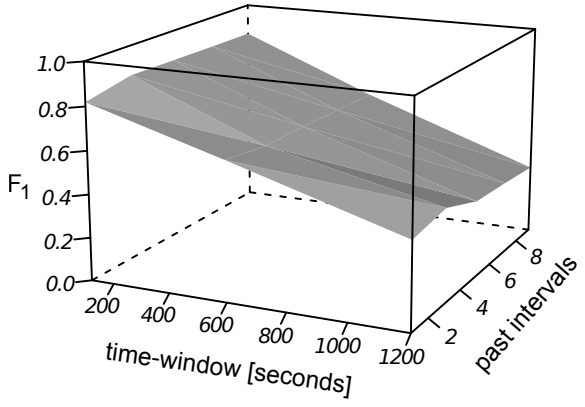
We adhere to a supervised learning procedure in our evaluation. Each mobile trace is split in two equal-sized temporal parts: the first period is used as the training set and the remaining part serves to report the prediction performance. We use a Support Vector Machine classification algorithm (using LIBSVM library [158]) under different parameter settings and used a validation set to avoid overfitting. We report the quality of the prediction using the F_1 score (also referred to as F-measure in the literature), expressed as the harmonic mean between precision ($\frac{TP}{TP+FP}$) and recall ($\frac{TP}{TP+FN}$) as defined by the confusion matrix (Table 5.5).

5.5.3 The effect of time-window duration and past data

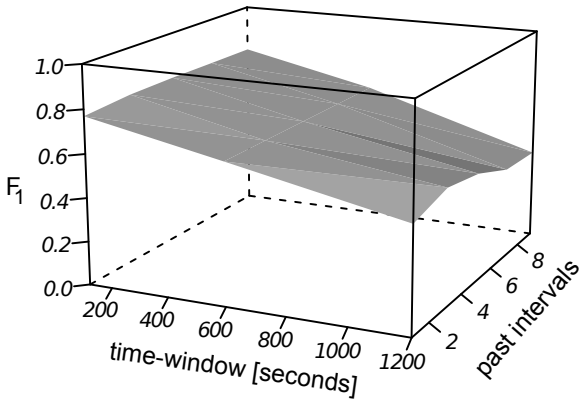
We analyze how the prediction performance is influenced by the duration of the time-window and the number of past intervals (time-windows) used in the prediction model. Aggregating data over longer durations may lose useful temporal information about the structure of the dynamic network. On the other hand, on more granular separation may capture important temporal patterns but also increase the computational cost.

We build prediction models that use $\{1, 3, 5, 7, 9\}$ time-windows and illustrate the results for the 1-contact case as we observed that the remarks made on this value are consistent with other κ values as well. For the size of the time-window we select the most granular duration (the probing interval used in each mobility trace) and two other values that represent $5\times$ and $10\times$ this duration. Thus, we consider time-windows of duration {120, 600, 1200} seconds for *Sig09* and *Infocom05* and use {15, 75, 150} seconds for *Rollernet* (which has a more granular probing rate).

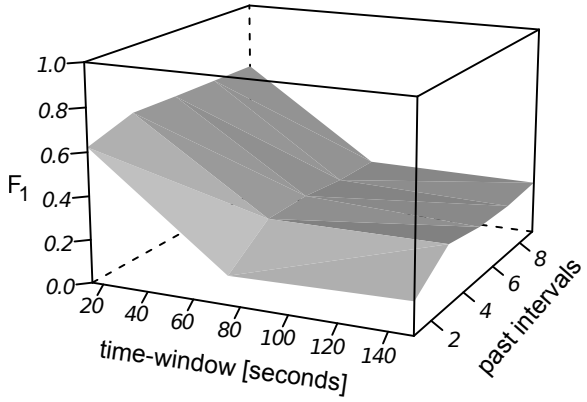
The results are presented in Figure 5.6 by means of 3D plots that represent the F_1 score as a function of the time-window duration and the number past intervals used in the prediction model. On the x-axis we examine different time-window durations and the y-axis (labeled *past intervals* in Figure 5.6) denotes the number of time-windows used in the prediction model. For example, a past interval of length 9 for a time-window of 1200 seconds means that, based on the contacts recorded during the previous 9 intervals of 1200 seconds, we predict contacts during the next 1200 seconds.



(a) *Infocom05*



(b) *Sig09*

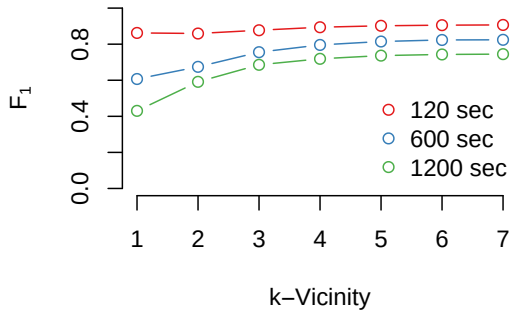


(c) *Rollernet*

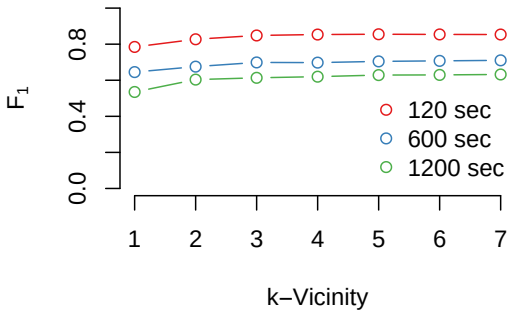
Figure 5.6: Prediction performance for different time-window durations and by varying the number of training time-windows (past intervals).

The figure illustrates that the most recent information plays the most important role in the prediction performance. For all three data sets, using data from the latest three time-windows achieves the highest performance and older information has little predictive power. This indicates that the most recent interactions are the most important in predicting the immediate future [1]. We can also observe that the longer the duration of the time-window, the less accurate the prediction performance. This suggests that aggregating data over longer durations is prone to larger errors. Taking the example of *Infocom05* (Figure 5.6(a)), the results show that predicting the contact opportunities during the next 2 minutes shows an F_1 score of 0.8 and the performance drops with 50% when trying to predict what will happen during the next 20 minutes. For *Rollernet*, which represents a more dynamic

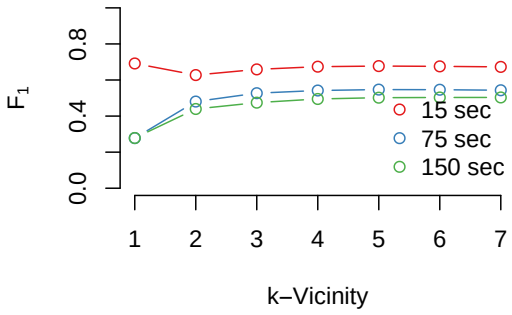
[1] These observations may apply only to these specific contact traces. For traces that span longer periods of time other periodicities could be observed.



(a) *Infocom05*



(b) *Sig09*



(c) *Rollernet*

Figure 5.7: The efficiency of predicting κ-contact relationships for different durations of the time-window. On the y-axis we represent the prediction performance and on the x-axis we vary the value of κ-contact from 1 to 7.

scenario, the drop of performance is even higher with a 70% decrease when trying to predict the contacts during the next 150 seconds compared to a 15-seconds time-window.

5.5.4 κ-contact prediction results

Based on the previous observations of the optimal number of past intervals (3 in our case) we assess the performance of predicting κ-contact relationships. We vary the value of κ from 1 to 7 and consider three durations for the time-window: {120, 600, 1200} seconds for *Infocom05* and *Sig09* and {15, 75, 150} seconds for *Rollernet*. The results are illustrated in Figure 5.7. First, we observe that predicting that two nodes will be in direct communication range shows particularly poor results for *Rollernet* data set, very dynamic

mobile settings, and when trying to make predictions over longer periods of time. Thus, in situations that involve important changes in the network topology, predicting that nodes will be in direct contact is prone to large errors.

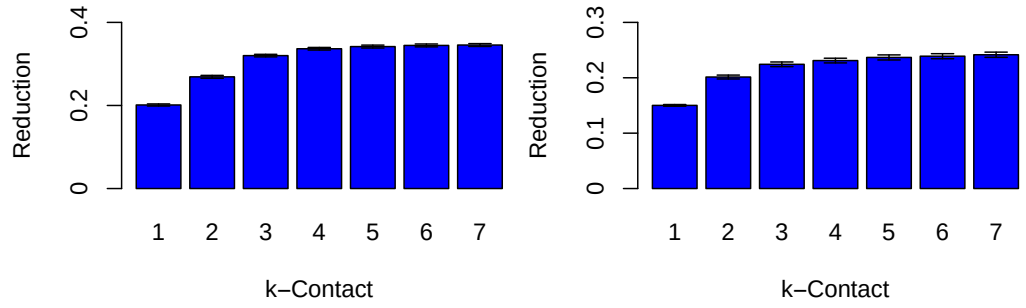
Relaxing the prediction objective beyond direct contact relationships reveals more accurate predictive power. Overall, the greater the value for κ the more effective the prediction performance. On average (for all mobility traces and different time-window durations) predicting that nodes will be at most at a distance 2, 3, and 4 shows an improvement of 7%, 10%, and 11% compared to the case where we want to predict direct encounters. While the improvement is important for small values of κ we notice that there is little benefit in extending the prediction for a κ greater than 3. The most significant increase, compared to the direct contact case, can be observed for $\kappa = 2$ with an average increase of 10% for *Rollernet*, 7% for *Infocom05*, and 6% for *Sig09*. The benefit is negligible when trying to predict the network change in the immediate horizon but it becomes significant when trying to make predictions over longer periods of time. Taking the case of *Infocom05* for a time-window of 1200 seconds and *Rollernet* for 150 seconds, predicting that nodes will be separated by at most two nodes (κ-contact = 3) reveals an improvement of 60% for *Infocom05* and 74% for *Rollernet* compared to the direct contact prediction case.

We provide two plausible explanations for these results. First, as we showed in Figure 5.4, a non-negligible number of nodes, although never in direct contact, they come at a 2-hop distance. By extending the prediction objective to 2-hop contacts, we include these potential events into consideration, which appear to have a more predictable nature. Then, as showed in Section 5.4 direct contacts between mobile users are scarce and short-lived, which makes them more difficult to predict in very dynamic scenarios and for longer time horizons. This explains the low prediction effectiveness observed with *Rollernet* and for longer time-windows for *Sig09* and *Infocom05*. Thus, extending the notion of contact to κ-contact gives us access to more stable connections (nodes leave direct connectivity but remains in κ-contact for longer durations) that reveal a more predictable nature.

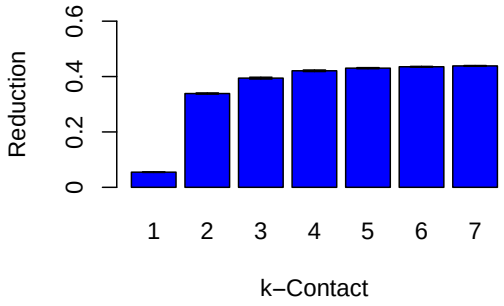
5.6 Practical implications

To capture the possible benefit that κ-contact prediction would bring in practical scenario we propose and evaluate the following use-case example.

We consider a content producer, located on the Internet, that regularly publishes content for a known group of collocated mobile users that communicate with the server using the cellular infrastructure. Content is categorized in topics. Users subscribe to these topics and content is pushed to users upon creation. We also consider that, in order to reduce the amount of cellular traffic caused by content delivery, the content producer collects data



(a) *Infocom05* (b) *Sig09*



(c) *Rollernet*

Figure 5.8: The percent of traffic with the infrastructure that can be reduced through κ-contact prediction and mobile opportunistic communications. On the y-axis we represent the traffic reduction compared to the case where content is sent to mobile users using only the infrastructure. On the x-axis we present different values for κ-contact.

about the contacts between mobile users and uses κ-contact prediction functionality to transfer the published objects to mobile users. More specifically, at the publication of a piece of content, instead of individually transmitting the content to each subscriber, the content producer optimizes the delivery process based on the predicted κ-contact opportunities. For example, if the server predicts that two users, interested in the same content, will be in κ-contact, a message is sent to only one of these nodes that will opportunistically forward the message to the peer node when they will be κ-contact. We also assume that nodes are capable of sensing their κ-vicinity and can detect when a targeted user is in κ-contact. To collect nearby topological knowledge, we assume the existence of a link-state protocol gathering nearby knowledge under the form of a connectivity graph. The implementation itself is beyond the scope of this study, yet previous analysis studied the impact of monitoring overhead [145].

We design the experimental setting using ONE simulation environment [159]. In our experiments we set the number of topics to 100. Each mobile node randomly subscribes to 20 up to 100 topics. For the prediction module, we use a time-window of 75 seconds for *Rollernet*, and 600 seconds for *Infocom05* and *Sig09*. Content is uniformly created throughout the duration of the experiments (that covers the duration of a mobility trace) and the results are averaged over 10 simulation runs. We also consider an infinite cache size at the user side and assume that one piece of content is small enough to fit into one message in the communication between content producer and the users and between the mobile users. To measure the impact of κ-contact prediction we report the reduction in the number of messages in the communication between the content producer and the mobile users when using κ-contact prediction module compared to a case where the content is individually sent to each user using the cellular infrastructure.

The results are presented in Figure 5.8. First, we observe that the greater the value of κ-contact, the greater the potential of traffic reduction. A significant improvement of predicting beyond direct neighbors is noticed for $\kappa = 2$, that shows an improvement of 6% for *Sig09*, 7% in *Infocom05*, and 30% for *Rollernet*. The potential traffic reduction is directly influenced by the characteristics of the connectivity traces: κ-vicinity properties (presented in Figure 5.4) and prediction performance (presented in Figure 5.7). Taking the example of *Sig09*, even if the effectiveness of the prediction showed little improvement for $\kappa = 2$ compared to $\kappa = 1$ the potential reduction is nevertheless important (6%). This is explained by the significant number of nodes located at a 2-hop distance and correctly predicted. The benefit is even more substantial in the case of *Rollernet*. By counting on the pairs of nodes connected at a 2-hop distance (that exceed the number of direct contact opportunities), the traffic reduction attains a performance of 33% compared to 5% when using only direct contact prediction. Thus, a κ-contact prediction model used in the context of mobile data offloading could be an effective solution considering that: by extending the pairwise vision from contact to κ-contact we consider more end-to-end transmission opportunities; and given that κ-contact interactions can be predicted more accurately than contact relationships.

5.7 Conclusions

In this chapter, we addressed the problem of predicting κ-contact opportunities between mobile users – predict if users will find themselves at a distance of at most κ-hops from one another. By analyzing three real-life contact traces, we observed that one can obtain better performances when predicting 2^+-contacts compared to the direct contact case. Using a supervised prediction framework, we studied the predictive nature of κ-contacts and

compared it with the traditional case of predicting contacts between nodes. Our results indicate that, in highly dynamic mobile settings (e.g., rollerblading scenario), predicting that nodes will remain at a distance of two hops from one another, can attain twice the performance of direct contact prediction. To assess the impact of these findings in a real-life application, we proposed a simulation experiment in which, by combining mobile opportunistic communications with κ-contact prediction one can reduce the amount of traffic used in the communication of mobile users with the infrastructure. Our results suggest that services that may benefit from contact predictions [160] can efficiently exploit the predictable nature of κ-contacts.

Chapter 6

Conclusions and future work

6.1 Summary

Mobile opportunistic networks provide a good solution to the challenging problem of mobile data offloading but the success of real-life deployments will depend on the capacity to better understand and predict mobile users' behavior.

In this dissertation we looked at new perspectives about user behavior that can be used to improve the efficiency of mobile opportunistic data offloading solutions. The first aspect proposed in this work is to study users' content access patterns, build models to predict content popularity, and adjust the availability of content based on the predicted users' demand. To better understand the difficulties and limitations of actually predicting web content popularity (in our case online news articles) we analyzed the popularity of articles published on two online news platforms. After studying the various prediction methods proposed in the literature, we analyzed the capacity of two of these methods to predict the popularity of news. Our results indicate that a linear model on a logarithmic scale is an effective solution to predict the popularity of online news. Furthermore, in the context of automatic online news ranking we showed that this method is also an effective solution to correctly rank news items by their future popularity; with a performance that can evenly match more customized learning to rank algorithms. To get real value out of these observations we showed that the ability to actually predict future content demand can improve the impact of proactive seeding used in the context of mobile opportunistic data offloading.

The second aspect addressed in this thesis is the capacity to predict κ-contact opportunities between mobile users – predict if users will find themselves at a distance of at most κ-hops from one another. By analyzing three real-life connectivity traces we observed that,

in a mobile scenario, one can obtain better performances in predicting that users will find themselves not necessarily in direct communication range but in the nearby vicinity separated by only few other mobile users. To assess the impact of these findings in a real-life application, we proposed a simulation scenario in which, by combining mobile opportunistic communications with κ-contact prediction, one can reduce the amount of traffic used in the communication of mobile users with the infrastructure. Our results suggest that services benefiting from contact predictions can efficiently exploit the predictable nature of κ-contacts.

6.2 Looking ahead

We identify several directions for the future work: (1) design more accurate content popularity prediction algorithms; (2) study the predictability of spatiotemporal connections in mobile opportunistic networks; and (3) design a real-life mobile data offloading engine that combines the capacity to learn and predict users' content demand and mobility patterns and uses it to better orchestrate the mobile data offloading decisions.

6.2.1 Improving the quality of the prediction

Even if research on predicting the popularity of web content has been an active area in the latest years there are many avenues that wait to be explored.

Predicting long-term popularity evolution. Most previous works addressed the problem of predicting the popularity of a web content up to a specific moment of time. While this is useful to detect in time future popular web items, a bigger impact would come from a long-term evolution forecast [36, 38]. Knowing this can provide important insights of how content progresses through different stages of popularity: initial growth, peak period, decline, and even popularity rebounds. Such information can help online advertisers or content delivery networks in making more profitable decisions, focusing on a web content during its peak period and wasting less resources on expired web items.

Building richer prediction models.In addition to early popularity measures, different studies have analyzed the predictive power of various features. We believe that this direction has not been fully explored and more work in finding more powerful predictive features would be valuable. For example, except for Bandari et al., which used the news source in their prediction model [65], to our knowledge no other work has studied the predictive power of content publisher. Yet, news columnists and video publishers attract a significant (and maybe predictable) audience on their own.

The topic of a web content plays, without doubt, an important role in its future popularity. The daily agenda of discussions in the Internet and mainstream media is centered

on major topics with limited and different life cycles. Thus, capturing trending topics and learning how to include them in prediction models can lead to a major breakthrough in prediction accuracy. Research in this field has made important advances in the recent years. Leskovec et al. found that the attention that online users pay to certain topics can accurately be described by six different time-series shapes [161]. In a novel approach to find trending topics on Twitter, Nikolov et al. proposed an algorithm that can detect trending topics earlier (with an average of 1.43 hours) than the internal algorithm used by Twitter [162].

Current work has showed that, for web items with very short lifecycle, timely predictions (within minutes after a web content has been posted) are a major challenge. News articles are a perfect example as they quickly become popular and "die-out" within hours. One way to improve the predictability of news would be to extract recurrent events over time, observe the level of interest that they generate, and predict when these future events will take place. Predicting global events in various fields (e.g., economy, seismology, society), as challenging as it may seem, is nevertheless plausible. Radinsky et al. have proposed two algorithms for this prediction task: PROFET, an algorithm that predicts the terms used in the future news based on the historical web query patterns [163]; and Pandit, a system that can predict future events given a certain news event [164].

Understanding and merging user activity, stemming from different web channels, is undeniably an important direction to follow. Up to now, Twitter feed has been used as the main source of information. But there are other rich opportunities to explore. For example, analyzing Web users' query behavior can unveil important insights about the popularity of certain topics, and the ability to predict search queries, as showed by Radinsky et al. [165], could be incorporated in a popularity prediction model. Wikipedia is also a valuable source of information. Important real-life events are quickly recorded on Wikipedia, and real-time monitoring of this channel can be transformed into valuable knowledge by a prediction model. Wikipedia Live Monitor is a good example of automatic monitoring tool that detects breaking news events by studying simultaneous user activity for certain topics edited in different languages [166, 167].

Beyond predictions. Studying online content popularity prediction should be not only useful in revealing new patterns in user dynamics, but also valuable in improving various web services. For instance, using the examples from the previous section, content producers (professionals or amateurs) can rely on the factors known to influence content popularity to build the genome of popular content. Although there are many factors that are difficult to control, creating content that is original (multiple copies of the same content has a negative impact on popularity [32]), fresh (the benefit of first-comer advantage [85]), emotional (stronger emotions are correlated to content viralness [79]), and by tagging it with popular

keywords (to appear in more popular recommendation lists [86]) can increase the likelihood of a web content to become popular. Then, online advertisers should try to figure out how to seize the opportunity of finding popular content in advance and design novel monetization strategies. Finally, there are few reports on how content popularity prediction can be used to design more effective networking solutions. Yet, predicting popularity dynamics can be used to design more scalable content delivery solutions (by proactively replicating content according the future demand) and to reduce the level of congestion caused by sudden bursts of content demands.

6.2.2 Smart proactive seeding

The proactive seeding strategy proposed in this work consists in prefetching popular content to randomly assigned mobile users to better cope with the future demand. Used as a real-life application this solution may seem primitive and rather unrealistic given that users may often need to spend additional resources to store content that is beyond their own interest. A reasonable alternative would be to predict individual user demand and adapt the preloading decisions accordingly. The benefit in this case is manyfold: mobile users will enjoy a better experience in accessing content on-the-go (e.g., reduced delay, better tolerance to network disconnections); content providers can better deliver information to mobile clients; and telecom operators can further improve the effect of proactive seeding (one content preload reduces the future network load with at least the same quantity).

One can go even further and imagine that in addition to predicting *what* users will consume the prediction engine could also reveal *when* the expected request will take place. The theoretical benefit of this approach is even greater as this additional knowledge can be used to build mechanisms for data traffic shaping [118].

Of course, creating a predictive behavior analytics engine is much more complex as it requires an additional amount of personal information about users' interest, usage patterns, or social behavior. But popular commercial applications such as Incoming TV, used for content recommendation and content delivery services, show that users are open to divulge more about their preferences to improve the quality of their mobile experience [168].

6.2.3 Predicting spatiotemporal contacts

As presented in Chapter 5, predicting future communications opportunities between mobile users can efficiently be used in the context of mobile data offloading. To make an even better use of this solution future work should try to extend the prediction objective to predicting when the κ-contact will take place and the duration of the connection. This information can then be used to decide when to initiate the transfer of a message or to postpone the transfer of bulky data if the duration of the communication opportunity is too short.

Predicting only instantaneous communication paths between mobile users gives access to only some of the data transfer opportunities in a mobile setting. In mobile opportunistic network environments, characterized by frequent disconnections, physical paths between users may never exist (or they may be too transient) but spatiotemporal paths can be more frequent. Formalized under the concept of temporal reachability graphs [169] this concept allows one to capture temporal communication capabilities between mobile users.

6.2.4 Mobile opportunistic data offloading engine

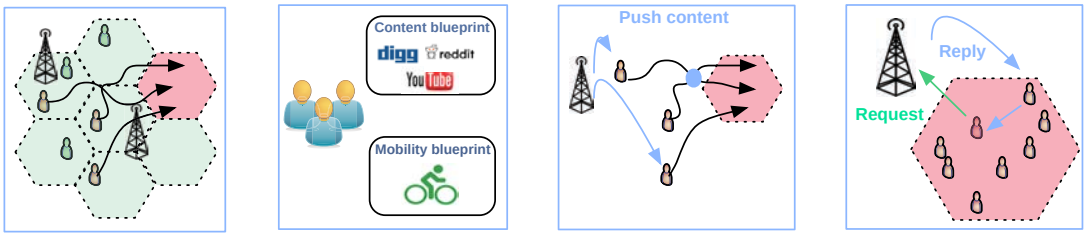


Figure 6.1: Opportunistic mobile data offloading procedure.

Finally, the solutions proposed in this work can be used as integrated components of an opportunistic mobile data offloading engine used by telecom operators to improve data traffic services in highly congested areas caused by large gatherings (e.g., concerts or sporting events). Concretely, this service, illustrated in Figure A.3, could be decomposed in the following steps:

- *Identify mobile participants*: for a predefined location that presents a high risk of being congested, identify the population of mobile users that will be within the area of interest.

- *Learning users' behavioral patterns*: for the population of users the prediction engine would need to track, understand, and identify patterns in user behavior. This includes learning data usage patterns (learn what users will likely consume inside congested areas) and mobility characteristics (learn the trajectory of users up to the location of interest).

- *Coordinated proactive seeding*: by predicting what users will consume inside the congested location the telecom operator can preload content during users' journey up to a specific location. This action can be performed using the cellular infrastructure

(when users pass thought areas with low data traffic), Wi-Fi access points, and device-to-device communications by predicting future encounters between mobile users.

- *Improving the service inside congested areas*: in addition to the proactive offloading strategies the telecom operator could also improve users' experience inside the congested area. In particular, during periods of congestion, due to an increased dropping probability, transferring data to certain mobile nodes can be difficult. By knowing which users will remain in the proximity of a target node (thought κ-contact prediction) the telecom operator may decide to use another mobile node as a proxy in the communication with the targeted mobile user.

Appendix A

Résumé en francais

A.1 Contexte et motivation

Les dernières années on été marqués par deux nouvelles tendances dans l'évolution de l'Internet. Tout d'abord, on a assisté à la démocratisation de la création du contenu. Stimulé par le progrès technologique apporté par les plateformes web 2.0 et la croissance continue des sites de réseaux sociaux, les internautes sont désormais capables de créer et de partager de contenu eux mêmes – et souvent être en concurrence avec les producteurs de contenu professionnels. Ensuite, nous assistons à un changement important vers un accès mobile à Internet. Équipés avec des dispositifs mobiles plus performantes (plus petites, moins chers, et avec plus des fonctionnalités) et sous une meilleure couverture mobile à haut débit, des utilisateurs expriment une grande envie pour consommer du contenu à tout moment du temps. Un bon exemple de la rapide prolifération des appareils mobiles est illustré en Figure 1.1 par les deux photos prises dans une intervalle de huit ans, et représentant des personnes réunis dans la Place Saint-Pierre.

Ces deux tendances ont stimulé les besoins des utilisateurs d'êtres connectés partout et à tout moment – à d'autres utilisateurs et à l'information – et de créer, consommer et partager du contenu dans un rythme sans précédent. Par exemple, chaque minute, les utilisateurs du monde entier envoient plus de 300 000 tweets [1] , partagent plus de 680 000 éléments de contenu sur Facebook [2], et téléchargent 100 heures de vidéo sur YouTube [3]. Et cette augmentation n'est pas une tendance isolée. Il s'avère que le volume global de données a augmenté à un taux de 50 % par an et il y a eu une augmentation de 40 fois par rapport à 2001 [4]. Et, même s'il y a des ressources pour stocker cette énorme quantité d'information (par exemple, ça coûte 600$ pour stocker toute la musique du monde [4]), mettre en place une infrastructure pour rendre le contenu toujours disponible reste un

objectif ambitieux.



(a) 2005 (b) 2013

Figure A.1: Personnes rassemblées dans la place Saint-Pierre dans une différence de huit ans (Source: NBC[2]).

Les opérateurs de télécommunications doivent faire face aux nouveaux défis en raison de la croissance de consommation de données mobiles. Traditionnellement, lorsque la capacité du réseau a atteint des limites critiques, les opérateurs comptaient sur certaines solutions techniques: acquérir du spectre additionnel, déployer des antenne-relais supplémentaires, ou de passer à la dernière technologie de communication mobile (par exemple LTE). Mais il est généralement estimé que ces solutions ne peuvent faire face aux défis des années à venir. Le spectre est une ressource limitée (et aussi de plus en plus chère), l'efficacité du spectre atteint vite ses limites, et l'installation de stations de base vient avec un coût important. En conséquence, de nouvelles solutions ont été proposées pour faire face à la consommation de données mobile prévue. Le délestage du trafic mobile de données est une solution attractive qui permet aux opérateurs de télécommunications de transférer une partie du trafic des réseaux cellulaires aux réseaux alternatifs à bas prix [5]. Dans ce contexte, les bornes Wi-Fi sont une ressource précieuse, car elles sont largement disponibles, elles ont un coût relativement bas, et un débit de données élève [6]. Les femto et picocellules sont aussi une alternative qui permettent une meilleure utilisation du spectre disponible [7]. Cependant, le potentiel de déchargement de ces solutions ne peut pas soutenir le taux de croissance de la consommation de données mobile et de nouvelles solutions sont nécessaires.

Les réseaux mobiles opportunistes ont été récemment proposés comme une solution attractive pour délestage du trafic de données mobile qui ne pose pas de contraintes de temps réel. Au lieu d'utiliser l'infrastructure de réseau cellulaire, les utilisateurs mobiles peuvent récupérer l'information à partir des autres utilisateurs mobiles qui sont dans la proximité et qui partagent un intérêt commun (et qui sont disponibles à partager l'information avec leurs voisins). Ce paradigme de communication ouvre de nouvelles perspectives sur la façon dont les utilisateurs mobiles peuvent générer et consommer du contenu tout le temps, mais

[2]http://instagram.com/p/W2FCksR9-e/

la conception des protocoles de communication pour les réseaux mobiles opportunistes est difficile (en raison de la mobilité des utilisateurs, il est difficile de faire des hypothèses sur l'existence d'un chemin entre les nœuds mobiles) et elle dépend en grande partie sur la capacité à comprendre le comportement des utilisateurs mobiles. *Ainsi, capturer le comportement des utilisateurs mobiles, découvrir des régularités, et construire des modèles de prédiction devient essentiel dans la conception des protocoles de communication dans les réseaux mobiles opportunistes.*

Jusqu'à présent, la plupart des études sur le comportement des utilisateurs mobiles ont été concentrées sur une meilleure compréhension de la mobilité humaine. Cela comprend des études sur la durée des contacts (et inter-contacts) entre les utilisateurs mobiles [8] et la périodicité des rencontres humaines [9], ou de comprendre les structures sociales sous-jacentes (physiques et en ligne) qui peuvent expliquer les tendances de la mobilité humaine [10, 11]. Étudier la mobilité humaine est essentielle, mais dans l'environnement complexe dans lequel les utilisateurs mobiles fonctionnent il y a d'autres aspects concernant le comportement des utilisateurs, tout aussi importants, qui peuvent être exploités.

L'objectif de cette thèse de doctorat est d'offrir des nouvelles perspectives sur la façon d'utiliser le comportement des utilisateurs mobiles dans la conception des solutions pour le délestage du trafic mobile. En particulier, on propose une nouvelle approche sur le problème de prédiction des contacts entre les utilisateurs mobiles et on met en avant l'idée que la conception des stratégies efficaces du délestage de données devrait non seulement prendre en considération la mobilité humaine, mais aussi l'information sur le trafic mobile de données consommées par les utilisateurs.

A.2 La problématique

Le scénario envisagé tout au long de ce travail, illustré dans la Figure 1.2, est composé de trois entités principales: un producteur de contenu, un opérateur de réseau de télécommunications, et un groupe d'utilisateurs mobiles qui se trouvent en proximité un de l'autre. Le producteur de contenu se trouve sur l'Internet et publie périodiquement du contenu pour le groupe des utilisateurs mobiles. L'opérateur de réseau de télécommunications fournit l'infrastructure pour la communication entre les utilisateurs mobiles et le producteur de contenu. Enfin, on considère un groupe des utilisateurs mobiles qui communiquent avec le producteur de contenu en utilisant l'infrastructure cellulaire et peut aussi communiquer directement entre eux en utilisant les techniques de communication directe d'appareil à appareil (par exemple, Bluetooth ou Wi-Fi Direct).

Ce scénario correspond à une certaine zone urbaine (par exemple, un campus universitaire ou un centre commercial) densément peuplée par des utilisateurs mobiles qui partagent

les mêmes intérêts dans aces au contenu. Dans ce contexte, on considère deux stratégies possibles pour les utilisateurs mobiles d'accéder au contenu. Dans l'approche classique, les utilisateurs s'appuient sur les services fournis par l'opérateur de télécommunications et récupérèrent individuellement le contenu en utilisant l'infrastructure. Mais, étant donné la corrélation temporelle et géographique d'accès au contenu, cette approche peut être considérée comme obsolète et inefficace, car il y a de fortes chances que le contenu pourrait être directement récupéré à partir des utilisateurs mobiles qui se trouvent en proximité. Une solution alternative sera d'utiliser les communications mobiles opportunistes (si la zone géographique est assez peuplée pour assurer de bons moyens de communication entre les utilisateurs) et de donner aux utilisateurs mobiles la possibilité de communiquer directement et de partager l'espace de stockage des autres utilisateurs mobiles voisins.

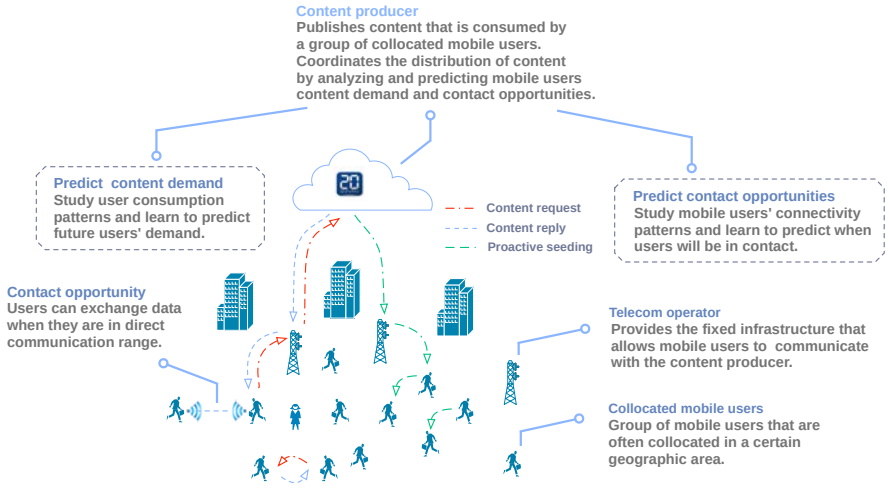


Figure A.2: Le scénario global considéré tout au long de ce travail est composé d'un producteur de contenu, situé sur l'Internet, un opérateur de télécommunications qui fournit l'infrastructure pour la communication entre les utilisateurs mobiles et le fournisseur de contenu, et un ensemble d'utilisateurs mobiles qui se trouvent en proximité.

Plusieurs solutions ont été proposées dans les dernières années pour le délestage de trafic de données, qui considèrent la coexistence de l'infrastructure avec les communications mobiles opportunistes pour décider quand et où (à quel utilisateur) de télécharger du contenu d'une manière préventive afin de réduire la communication des utilisateurs mobiles avec l'infrastructure [12,13]. Mais les solutions actuelles des communications opportunistes sont assez myopes dans le sens où ils ne bénéficient pas pleinement de la connaissance sur le comportement de l'utilisateur. Par exemple, en observant comment les utilisateurs consument du contenu, on peut imaginer un modèle capable de prédire la demande future et

adapter le fonctionnement du réseau en fonction de ce besoin. En outre, de plus grands avantages peuvent être atteints par le suivi de la mobilité humaine.

Le problème traité dans cette thèse est de proposer des nouvelles solutions pour les communications mobiles opportunistes, des solutions fondées sur une connaissance globale sur la demande de contenu et la connectivité humaine. En particulier, on répond à deux problèmes principaux:

• **Problème 1**

Quel contenu à envoyer? Étant donné la grande quantité de contenu publié quotidiennement, la répartition inégale de l'intérêt des utilisateurs, et la popularité non stationnaire du contenu (la popularité d'un morceau de contenu évolue au fil du temps) il est important de décider quel contenu à télécharger aux utilisateurs et le nombre des copies pour mieux gérer la future demande des utilisateurs. Cela permet de construire des techniques de dissémination préventive qui sont adaptées à l'évolution dynamique de la popularité du contenu.

Stratégie. Tout d'abord, on veut comprendre dans quelle mesure la popularité du contenu web peut-être prédite. On prend comme exemple le contenu publié sur un journal en ligne et on étudie la popularité des articles publiés sur deux plates-formes web. On étudie les différentes méthodes de prédiction qui ont été proposées dans la littérature, on choisit ceux qui sont adaptées à ce type de contenu, et on étudie leur capacité à prédire la popularité des articles. On évalue ensuite l'impact de ces résultats pour une solution de délestage de données mobile en utilisant les communications mobiles opportunistes.

• **Problème 2**

Comment choisir les utilisateurs pour télécharger les données mobiles? Étant donné la nature opportuniste des rencontres humaines qui sont, dans une certaine mesure, prévisibles il faut décider sur la façon de mieux organiser le processus de téléchargement du contenu en choisissant où (à quels utilisateurs) pour envoyer le contenu.

Stratégie. On analyse des différentes traces des contacts humains et on étudie la prévisibilité des contacts humains. Étant donné la nature assez imprévisible de ces relations on étend l'analyse au cas κcontact – prédire si les utilisateurs vont se trouver à une distance d'au plus κutilisateurs. Pour évaluer l'impact de ces résultats dans une application réelle, on propose une expérience de simulation dans laquelle, en combinant les communications mobiles opportunistes avec la prédiction κcontact, on peut réduire la quantité de trafic utilisé dans la communication de nœuds mobiles avec l'infrastructure.

A.3 Contributions de cette thèse

A.3.1 Une synthèse sur les algorithmes de prédiction de la popularité du contenu web.

Lorsqu'on étudie la popularité du contenu web, il n'est pas clairement établi qu'il existe un modèle de prédiction qui pourrait être appliqué à tous les scénarios possibles, ni que la création d'un modèle de prédiction générique est un objectif réalisable. Les raisons sont que les résultats de prévisions sont influencés par le type de contenu en ligne, le cadre du site, et la disponibilité de l'information prédictive. Ainsi, dans le domaine des médias sociaux plusieurs méthodes de prédiction de la popularité ont été proposées et évaluées pour différents types de contenu web.

La première contribution de cette thèse est une synthèse sur les méthodes de la prédiction de popularité de contenu web. Ce sujet de recherche est devenu un domaine actif et un grand nombre de méthodes de prédiction pour différents types de contenus web ont été proposées dans les dernières années. Pour structurer les méthodes de prédiction existantes, on propose une classification fondée sur le type d'information utilisée dans la prédiction. On présente les performances des différentes méthodes de prédiction, on expose les caractéristiques qui ont fait preuve de bonnes capacités prédictives, et on révèle les facteurs connus pour influencer la popularité du contenu web.

Dans le monde numérique, le contenu web est devenu l'attraction principale. Que ce soient des informations utiles, du divertissement pour les utilisateurs d'Internet, ou une possibilité des affaires pour les entreprises de marketing et les fournisseurs de contenu le contenu web est un atout sur Internet. Dans le même temps, la croissance dans l'innovation des médias sociaux, la facilité de la création du contenu et les coûts de publication faibles, ont créé un monde saturé d'information. Pourtant, l'écosystème en ligne adhère à une société "winner-take-all": l'attention est concentrée sur quelques pièces de contenu alors que la majorité reste inconnue. Dans ce contexte, trouver le contenu web qui sera populaire devient de la plus haute importance. Les utilisateurs en ligne, inondés par information, peuvent réduire l'encombrement et concentrer leur attention – la ressource la plus précieuse dans le monde d'Internet – sur l'information la plus pertinente. Dans un monde où les entreprises dépensent jusqu'à 30 % de leur budget dans le marketing en ligne [14], repérer le plus vite possible la prochaine étoile montante de l'Internet peuvent maximiser leurs revenus grâce à un meilleure placement de la publicité. En outre, étant donné la croissance de trafic Internet, les réseaux de distribution de contenu peuvent s'appuyer sur des méthodes de prédiction de la popularité et d'allouer les ressources d'une façon proactive en fonction de futures demandes des utilisateurs.

Le terme *de contenu web* est effectivement générique et il définit d'une façon générale

tout type d'information sur un site web. Il peut faire référence à la fois à l'information transmisse et de l'objet individuel utilisé pour transmettre l'information. Dans ce travail, on définit le contenu web comme tout élément individuel à la disposition du public sur un site web et qui contient une mesure qui reflète un certain intérêt manifesté par une communauté en ligne.

La notion de la popularité du contenu web est subtile, au-delà du nombre habituel des pages vues. Avant le web 2.0, la mesure la plus importante était le nombre des fois où une page web est affichée, mais maintenant, avec la prévalence de plates-formes de médias sociaux, il y a de nouveaux indicateurs qui reflètent l'intérêt des utilisateurs. En réponse à la publication de contenu, les utilisateurs peuvent désormais exprimer leur opinion face au contenu par commentaires et évaluations, ou partager davantage dans leurs cercles sociaux en ligne (par exemple, Facebook, Twitter, ou Digg). Ces mesures alternatives capturent l'engagement plus profond des utilisateurs et fournissent de précieuses informations complémentaires au nombre de pages vues: les évaluations améliorent la qualité des publications, les commentaires augmentent le temps passé sur une page web, et le partage sur les réseaux sociaux donne au contenu une plus grande notoriété. En outre, ces mesures capturent des habitudes différentes comme les utilisateurs ont des préférences différentes (e.g., commenter, évaluer, ou partager) [15–18]. Dans ce contexte, l'étude de ces paramètres individuellement ou comment ils se rapportent les uns aux autres [19, 20] offre une perspective plus complète de ce que signifie réellement la popularité de contenu.

La prédiction de la popularité du contenu web est une tâche difficile. Tout d'abord, des différents facteurs connus pour influencer la popularité du contenu, telles que la qualité du contenu ou la pertinence pour les utilisateurs, sont difficiles à mesurer. Ensuite, d'autres facteurs, tels que la relation entre des événements dans le monde physique et le contenu sont difficiles à capturer et utiliser dans un modèle de prédiction. En outre, au niveau microscopique, l'évolution de la popularité du contenu peut être décrite par des interactions en ligne complexes et cascades d'information, qui sont difficiles à prédire [21–23].

La prédiction de la popularité du contenu web est devenue un domaine de recherche actif et un grand nombre de méthodes de prédiction pour différents types de contenu web ont été proposées dans les dernières années. Dans une première étape, pour avoir une meilleure compréhension sur les enjeux et les solutions existantes, dans ce chapitre, on examine d'un point de vue général le problème de la prédiction de la popularité du contenu web. On passe en revue l'état actuel de la recherche dans ce domaine, on propose une classification qui nous permet de structurer les différentes méthodes de prédiction, et on décrit brièvement les principales méthodes de prédiction.

A.3.2 Prédire la popularité des articles

Pour mieux comprendre le problème de la prédiction de la popularité du contenu web, étant donnée la panoplie des méthodes appliquées aux différents types de contenu sur Internet, on étudie la capacité de prédire la popularité du contenu web en utilisant des traces web réels. On choisit comme contenu en ligne des articles de presse, un type de contenu qui saisit l'attention d'un nombre important des utilisateurs. Il s'agit d'un type de contenu qui peut facilement être produit, avec une petite taille et un faible coût de production – des propriétés qui le rend intéressant pour être massivement propagé dans les plateformes sociales en ligne et particulièrement apprécié par les utilisateurs mobiles.

On étudie deux plateformes des articles en ligne pour mieux comprendre comment les articles sont publiés par les agences de news et consommés par les lecteurs et on analyse la capacité de prédire la popularité des articles. On se concentre sur une dimension de la popularité du contenu et on considère le *nombre des commentaires* comme un évaluateur implicite de l'intérêt suscité par les articles. On évalue la performance de deux méthodes de prédiction, qui sont adaptées pour le type d'informations contenues dans notre ensemble de données, et on étudie leur capacité à prédire la popularité des articles en ligne.

En plus de prédire la valeur exacte de l'attention qu'un contenu va générer, dans une autre situation pratique, il peut être utile de classer les articles en fonction de leur popularité future. Comme l'intérêt des utilisateurs en ligne pour le contenu web est souvent inégal (avec des objets populaires étant extrêmement populaires) trouver les objets les plus populaires est souvent une solution assez bonne pour les applications qui bénéficient de prédire les préférences des utilisateurs. Par exemple, une approche qui pré-télécharge du contenu s'avère une solution robuste pour anticiper les demandes futures de contenu [89]. Ainsi, on étudie l'efficacité d'utiliser des méthodes de prédiction pour classer les articles en fonction de leur popularité future et de les comparer avec différentes heuristiques et une méthode d'apprentissage plus personnalisée de classification.

A.3.3 Pré-téléchargement du contenu fondé sur la prédiction de la popularité du contenu

Pour bénéficier de la prédiction de la popularité du contenu web, on étudie l'effet de cette solution dans le cadre du délestage du trafic mobile de données. En particulier, on propose la conception d'une stratégie proactive de délestage de données combiné avec les communications mobiles opportunistes qui peuvent aider les opérateurs de télécommunications à réduire le trafic de données pendant les périodes de charge.

Il existe différentes stratégies utilisées par les opérateurs de télécommunications pour faire face à la consommation croissante de trafic mobile de données. Les actions typiques sont d'optimiser la capacité du réseau (grâce à une meilleure planification du trafic), de

mettre à jour la technologie de réseau de prochaine génération (par exemple, LTE), ou d'acheter des blocs supplémentaires de spectre. Les alternatives plus récentes – moins chères et plus facile à déployer – sont construites sur la notion de délestage de données: l'utilisation des réseaux complémentaires pour déplacer dans le temps et dans l'espace le trafic de données qui, est à l'origine, destiné à traverser l'infrastructure cellulaire.

Les réseaux mobiles opportunistes offrent une bonne alternative pour décharger le trafic mobile de données qui ne posent pas de contraintes de temps réel. En permettant aux utilisateurs mobiles d'accéder à l'espace de stockage des utilisateurs qui se trouvent implantés au même endroit, les demandes de contenu peuvent être traitées par des communications opportunistes et donc de réduire le trafic de données ciblées à traverser l'infrastructure cellulaire [12]. Le téléchargement proactif du contenu (pré-téléchargement du contenu dans les appareils des utilisateurs mobiles avant la demande du contenu) a souvent été utilisé dans le contexte des communications opportunistes mobiles, où, pour réduire l'effet d'une faible connectivité de réseau, le contenu est pré-téléchargé dans l'espace de la mémoire cache de certains utilisateurs mobiles qui peuvent desservir les demandes futures du contenu [117]. Mais cette stratégie peut également être utile pour le délestage de données mobile, où, en anticipant la demande future d'un utilisateur, le contenu peut-être pré-téléchargé pendant les périodes de faible trafic de données afin de réduire la quantité de trafic à des instants futurs de temps [118].

L'avantage de pre-téléchargement dépend de la capacité d'anticiper les demandes futures des utilisateurs. Auparavant, lorsque des solutions similaires ont été utilisés pour réduire l'effet des goulot d'étranglement du réseau, prédire la demande de contenu des utilisateurs a été considérée comme une tâche difficile et des heuristiques simples ont été proposées pour détecter les futurs objets web les plus populaires [89]. Des découvertes récentes dans le domaine des médias sociaux montrent que la popularité du contenu web peut être prédite, un résultat qui peux améliorer l'impact de téléchargement proactif. On construit sur les résultats de ces constatations et on étudie l'effet de l'utilisation d'une méthode de prédiction de popularité réelle en tant que composant intégré de la décision de téléchargement proactif.

A.3.4 Prédire les événements κ-contact entre les utilisateurs mobiles

On étudie le problème de la prédiction des connectivités entre les utilisateurs mobiles. Étant donné la capacité assez limitée de prédire les contacts entre les utilisateurs mobiles, on étend le cadre de la prédiction pour le cas κ-contact – prédire si les utilisateurs mobiles vont se trouver à une distance d'au plus κnœuds un de l'autre. En utilisant un cadre de prédiction supervisé on analyse la prédictibilité de κ-contacts sur trois traces de contacts de la vie réelle et on observe qu'on peut atteindre de meilleures performances si l'on veut prédire que les utilisateurs ne seront pas en contact direct, mais dans à proximité. Pour

évaluer l'impact de ces conclusions dans un déploiement réel, on propose une expérience de simulation dans laquelle, en combinant les communications opportunistes mobiles avec un module de prédiction κ-contact, on peut réduire le trafic utilisé dans la communication de nœuds mobile avec l'infrastructure. Ces résultats indiquent qu'on peut efficacement exploiter la nature la plus prévisible de κ-contacts pour créer des meilleurs services de communications mobiles.

La conception des protocoles de communication dans les réseaux mobiles opportunistes dépend en grande partie sur la capacité à comprendre les caractéristiques de la mobilité humaine. Au cours des dernières années, plusieurs études ont révélé des résultats importants sur la durée des contacts et inter-contact entre les utilisateurs mobiles [8, 136, 137], la périodicité de ces rencontres [9], et les structures créées par les interactions humaines [10, 138–140]. Dans le contexte des réseaux mobiles opportunistes, les caractéristiques de la mobilité peuvent ensuite être utilisées pour concevoir des mesures qui facilitent la prédiction des interactions entre les utilisateurs mobiles. Cela consiste à l'utilisation de la fréquence des contacts pour identifier les similitudes entre les caractéristiques de la mobilité [141] ou de trouver les utilisateurs mobiles fortement connectés qui pourraient servir comme porteurs des messages [142]. Bien que ces mesures puissent servir comme heuristiques pour prédire les contacts entre les utilisateurs mobiles, ils ont une capacité limitée pour détecter les prochains contacts humains. Une approche plus avantageuse de ce problème, mais aussi plus laborieux, est de créer un modèle capable de prédire les contacts entre les utilisateurs mobiles.

Des études récentes ont traité cette question de la prédiction de contacts – prédire si deux nœuds vont être dans un rayon de transmission directe – et ont révélé que, avec une bonne méthode de prédiction et des caractéristiques prédictives, les contacts entre les utilisateurs mobiles sont, dans une certaine mesure, prévisibles [143]. Ce résultat est important, car il permet en fait de prédire les rencontres humaines et peut servir pour développer des protocoles de communication plus efficaces.

Mais les relations entre les paires des utilisateurs mobiles peuvent être décrites en dehors de la vue binaire (contact / inter-contact) comme souvent les individus peuvent se trouver hors de portée de transmission directe, mais toujours à proximité. Ainsi, pour avoir une vue plus complète sur les possibilités de communication disponibles, la notion élargie de contact, définit comme κ-contact, a été récemment proposée [145]. Les analyses précédentes ont montré que, étant donné que les contacts entre les utilisateurs mobiles offrent une compréhension biaisée et sous optimale du réseau de communication, les études sur κ-contacts permettent d'avoir une compréhension plus complète des possibilités de

Les propriétés et l'impact du κ-voisinage dans les réseaux mobiles opportunistes ont été étudiés par Phe-Neau [144].

communication de bout en bout.

Dans ce chapitre, on fournit de nouvelles informations sur les relations κ-contact et on montre qu'étudier seulement les contacts directs offre une vue limitée sur les possibilités de communication entre les paires des nœuds. On étudie ensuite la prédictibilité de κ-contacts. En utilisant des données provenant de trois traces des contacts humains, on compare la capacité de prédire les κ-contacts avec le cas classique de prédiction des contacts directs entre les utilisateurs mobiles. On montre que les relations κ-contacts sont plus prédictibles que les relations de contact direct. Pour mesurer l'impact de ces résultats dans une application réelle, on analyse l'impact d'une solution qui utilise un modèle de prédiction κ-contact pour le délestage de données mobile. Par simulation, on montre qu'il existe un grand potentiel de s'appuyer sur la prédiction κcontact par rapport au cas de contact direct.

A.4 Conclusions

Les réseaux opportunistes mobiles offrent une bonne solution au problème difficile de délestage de trafic mobile de données, mais le succès des déploiements réels dépend de la capacité de mieux comprendre et de prédire le comportement des utilisateurs mobiles.

Dans cette thèse, on a étudié des nouvelles perspectives sur le comportement des utilisateurs qui peuvent être utilisés pour améliorer l'efficacité des solutions de délestage de données mobiles. Le premier aspect proposé dans ce travail consiste à étudier l'accès au contenu, de construire des modèles pour prédire la popularité du contenu et d'ajuster la disponibilité de contenu basé sur la demande prédite des utilisateurs. Pour mieux comprendre les difficultés et les limites de la prédiction de popularité de contenu web (dans notre cas les articles de presse), on a analysé la popularité des articles publiés sur deux plates-formes d'information en ligne. Après avoir étudié les différentes méthodes de prédiction proposées dans la littérature, on a analysé la capacité des deux de ces méthodes pour prédire la popularité des articles de presse. Les résultats indiquent qu'un modèle linéaire sur une échelle logarithmique est une solution efficace pour prédire la popularité des articles. En outre, on a montré que ce modèle de prédiction est aussi une solution efficace dans le contexte de classification des articles basés sur leur popularité; avec une performance qui peut être assez bonne que les méthodes de classification automatique. Pour bénéficier de ces observations, on a montré que la capacité de prédire la demande future du contenu peut améliorer l'impact de pré-téléchargement utilisé dans le contexte de délestage de données mobile.

Le deuxième aspect traité dans cette thèse est la capacité de prédire les κ-contacts entre les utilisateurs mobiles – prédire si les utilisateurs vont se trouver à une distance d'au plus κnœuds un de l'autre. En analysant trois traces de connectivité de la vie réelle, on a ob-

servé que, dans un scénario mobile, on peut obtenir de meilleures performances si on prédit que les utilisateurs vont se retrouver pas nécessairement dans un rayon de communication directe, mais toujours dans la proximité, séparés par seulement quelques autres utilisateurs mobiles. Pour évaluer l'impact de ces résultats dans une application réelle, on a proposé un scénario de simulation dans lequel, en combinant les communications mobiles opportunistes avec la prédiction de κcontact, on peut réduire la quantité de trafic utilisé dans la communication des utilisateurs mobiles avec l'infrastructure cellulaire. Nos résultats indiquent que les services bénéficiant de prévisions de contacts peuvent efficacement exploiter la nature prévisible de κ-contacts.

A.5 Perspectives

On propose plusieurs directions pour les travaux futurs: (1) trouver des algorithmes plus efficaces pour prédire la popularité de contenu; (2) étudier la prédictibilité des connexions spatio-temporelles entre les utilisateurs dans les réseaux mobiles opportunistes; (3) et créer un moteur de délestage de données qui combine la capacité d'apprendre et de prédire l'accès au contenu des utilisateurs et leur mobilité et qui utilise ces informations pour mieux orchestrer les décisions du délestage de trafic mobile.

A.5.1 Améliorer la qualité de la prédiction de popularité

Même si dans dernières années, il y a eu beaucoup de découvertes sur la capacité de prédire la popularité du contenu, il reste encore des directions importantes qui doivent être explorées.

Prédire l'évolution du popularité de contenu. La plupart des travaux antérieurs s'occupent du problème de la prédiction de popularité du contenu web à un moment précis du temps. Bien que cette approche permette de détecter très vite les articles web qui vont devenir très populaires, un plus grand impact sera de prédire l'évolution de la popularité à long terme [36, 38]. Cette stratégie peut fournir des indications importantes sur la façon dont le contenu évolue à travers des différentes étapes de la popularité: la croissance initiale, la période de pointe, le déclin, et même les rebonds de popularité. Cette information peut aider la publicité en ligne et les réseaux de diffusion de contenu à prendre des décisions plus rentables, en mettant l'accent sur le contenu web lors de sa période de pointe et de gaspiller moins de ressources sur les articles web qui ne présentent plus d'intérêt.

Créer des modèles de prédiction plus riches. La plupart des modèles de prédiction utilisent seulement la popularité du contenu juste après la publication – et ignorent les autres caractéristiques du contenu – pour prédire la popularité finale. Une direction importante pour les travaux futurs, qui n'a pas été suffisamment explorée, ce sera de trouver

des nouvelles caractéristiques prédictives. Par exemple, excepter Bandari et al., qui utilisent l'éditeur des articles en ligne dans un modèle de prédiction [65], aucun autre travail n'a pas étudié le pouvoir prédictif de l'éditeur de contenu. Pourtant, les chroniqueurs de presse et les éditeurs vidéo attirent un public important – et peut-être prévisible.

Le sujet du contenu web joue, sans doute, un rôle important dans sa popularité. Les discussions sur l'Internet et les médias traditionnels sont centrées sur des sujets avec des cycles de vie limités et différents. Ainsi, capturer les nouvelles tendances et apprendre comment les inclure dans les modèles de prédiction peuvent conduire à des découvertes capitales pour améliorer la précision de la prédiction. La recherche dans ce domaine a fait des progrès importants dans les dernières années. Leskovec et al. ont constaté que l'attention que les utilisateurs en ligne donnent aux certains sujets peut-être décrite avec précision par six formes de séries chronologiques différents [161]. Dans une nouvelle approche, pour trouver les sujets populaires sur Twitter, Nikolov et al. ont proposé un algorithme qui permet de détecter les sujets populaires plus tôt (avec une moyenne de 1,43 heures) que l'algorithme privé de Twitter [162].

Les travaux actuels ont montré que, pour les articles web avec très peu cycle de vie, les prévisions rapides (quelques minutes après le contenu web a été publié) représentent un défi majeur. Les articles de presse sont un bon exemple, car ils deviennent populaires très vite et perdent leur intérêt dans quelques heures. Une façon d'améliorer la predictibilité de popularité des articles serait d'extraire des événements récurrents au fil du temps, observer le niveau d'intérêt qu'ils génèrent, et prédire quand ces événements futurs auront lieu. Prédire les événements majeurs dans différents domaines (par exemple, l'économie, la sismologie, la société), aussi difficile que cela puisse paraître, est néanmoins plausible. Radinsky et al. ont proposé deux algorithmes pour ce type de prédiction: PROFET, un algorithme qui prédit les termes utilisés à l'avenir dans des articles de presse basés sur historiques des requêtes web [163]; et Pandit, un système qui peut prédire des événements futurs à partir d'un certain événement existant [164].

Au-delà des prévisions. Prédire la popularité du contenu en ligne devrait être non seulement utile pour trouver des nouvelles tendances dans la dynamique des utilisateurs, mais aussi précieux pour améliorer certains services web. Par exemple, en utilisant les résultats de la section précédente, les éditeurs des contenus (professionnels ou amateurs) peuvent s'appuyer sur les facteurs connus d'influencer la popularité du contenu pour construire le génome de contenu populaire. Bien qu'il existe des nombreux facteurs qui sont difficiles à contrôler, créer de contenu qui est original (copies multiples du même contenu ont un impact négatif sur la popularité [32]), nouveau (l'avantage de la première apparition [85]), émotionnel (les émotions fortes sont corrélées au partage en ligne du contenu [79]), et par le marquage électronique avec des mots-clés populaires (pour paraître dans des listes de

recommandations plus populaires [86]) peut augmenter la probabilité du contenu web de devenir populaire. Ensuite, la publicité en ligne devrait essayer de comprendre comment saisir l'occasion de trouver du contenu populaire à l'avance et de créer des stratégies de monétisation. Enfin, il existe peu de rapports sur la façon dont la prédiction de contenu peut être utilisé pour créer des meilleures solutions pour les réseaux de distribution de contenu. Pourtant, la prédiction dynamique de la popularité peut être utilisée pour créer des solutions de distribution de contenus plus extensibles (répliquer le contenu d'une façon proactive, selon la demande future) et de réduire le niveau de congestion causée par des requêtes soudaines de demandes de contenu.

A.5.2 Pré-téléchargement intelligent

La stratégie de pré-téléchargement proposée dans ce travail consiste à récupérer à l'avance le contenu populaire et de l'assigner au hasard aux utilisateurs mobiles pour mieux faire face à la demande future. Utilisée dans une situation réelle, cette solution peut sembler primitive et peu réaliste étant donné que les utilisateurs auront besoin de consacrer des ressources supplémentaires pour stocker du contenu qui peut être au-delà de leur propre intérêt. Une solution raisonnable serait de prévoir la demande individuelle des utilisateurs et d'adapter les décisions de pré-téléchargement en conséquence. L'avantage dans ce cas est multiple: les utilisateurs mobiles peuvent profiter d'une meilleure expérience dans leur accès au contenu (par exemple, un délai réduit, une meilleure tolérance aux déconnexions du réseau); les fournisseurs de contenu peuvent mieux fournir des informations aux clients mobiles; et les opérateurs de télécommunications peuvent améliorer l'effet de pré-téléchargement (le pré-téléchargement du contenu réduit dans l'avenir la charge du réseau avec au moins la même quantité).

On peut même aller plus loin et imaginer que, en plus de prédire ce que les utilisateurs consomment, le moteur de prédiction pourrait aussi révéler le moment lorsque la demande attendue aura lieu. L'avantage de cette approche est encore plus grand comme cette connaissance supplémentaire peut être utilisée pour construire des mécanismes de mise en forme du trafic de données [118].

La création d'un moteur d'analyse de comportement humain est beaucoup plus complexe, car elle nécessite des renseignements personnels supplémentaires sur l'intérêt des utilisateurs, les habitudes d'utilisation, et le comportement social. Mais des applications commerciales populaires comme la vidéo mobile, utilisées pour la recommandation du contenu et les services de diffusion de contenu, montrent que les utilisateurs sont ouverts à divulguer plus d'informations sur leurs préférences pour améliorer la qualité de leur expérience mobile [168].

A.5.3 Prédire les contacts spatio-temporelles

On a montré que prédire les possibilités de communication entre les utilisateurs mobiles peut être utile dans le contexte de la téléphonie mobile (pour le délestage des données mobiles). Pour améliorer l'impact de cette stratégie, les travaux futurs devraient essayer d'étendre l'objectif de la prédiction et de prédire le moment quand le κ-contact aura lieu (et la durée de la connexion). Cette information peut ensuite être utilisée pour décider le moment pour lancer le transfert d'un message ou de retarder le transfert des données volumineuses si la durée de la connexion sera trop courte.

Prédire les connexions instantanées entre les utilisateurs mobiles offre des possibilités limitées pour le transfert de données dans le cadre de communications mobiles opportunistes. Dans ce scénario, caractérisé par des déconnexions fréquentes, les chemins physiques entre les utilisateurs peuvent être inexistants (ou ils peuvent être transitoires) mais des chemins spatio-temporels peuvent être plus fréquents. Formalisé sous le concept de graphes d'accessibilité temporels [169] cette stratégie permet de capturer les communications temporelles entre les utilisateurs mobiles.

A.5.4 Moteur pour le délestage opportuniste de trafic mobile de données

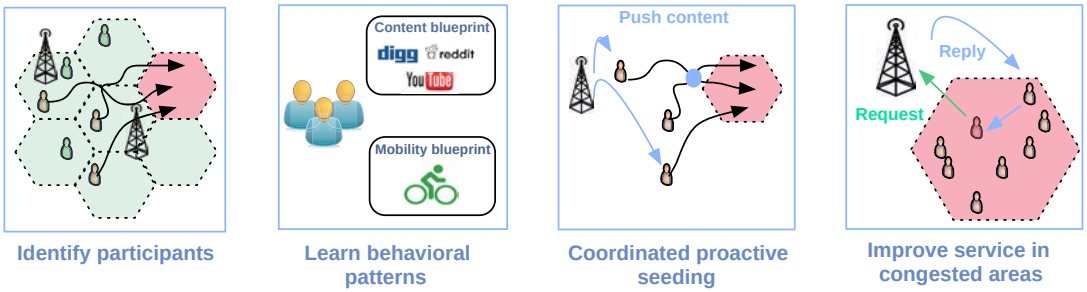


Figure A.3: La procédure utilisée dans cette thèse pour le délestage opportuniste de trafic mobile de données.

Enfin, les solutions proposées dans ce travail peuvent être utilisées comme composantes intégrés pour un moteur de délestage de trafic mobile de données utilisé par les opérateurs de télécommunications pour améliorer les services de transfert de données dans les zones encombrées causés par les grands rassemblements (par exemple, des concerts ou des événements sportifs). Concrètement, ce service, illustré dans la Figure 1.3, pourrait être décomposé dans les étapes suivantes:

- *Identifier les participants mobiles*: pour un endroit prédéfini, qui présente un risque

élevé d'être encombré, il faut identifier la population d'utilisateurs mobiles qui seront dans la zone d'intérêt.

- *Apprendre les comportements des utilisateurs*: pour la population d'utilisateurs, le moteur de prédiction aurait besoin de suivre, comprendre, et identifier les tendances dans le comportement des utilisateurs. Cela comprend l'apprentissage des habitudes d'utilisation des données (savoir ce que les utilisateurs vont consommer à l'intérieur des zones congestionnées) et les caractéristiques de la mobilité (savoir la trajectoire des utilisateurs dans la zone d'intérêt).

- *Le pré-téléchargement coordonné*: savoir ce que les utilisateurs vont consommer à l'intérieur du lieu encombré, l'opérateur de télécommunications peut télécharger du contenu pendant les voyages des utilisateurs à un emplacement spécifique. Cette action peut être effectuée en utilisant l'infrastructure cellulaire (lorsque les utilisateurs se déplacent à travers des zones avec une bonne connectivité cellulaire), les points d'accès Wi-Fi, et la communication de dispositif à dispositif avec la prédiction des rencontres entre les utilisateurs mobiles.

- *Améliorer le service à l'intérieur des zones encombrées*: en plus des stratégies de pré-téléchargement, l'opérateur de télécommunications pourrait également améliorer l'expérience des utilisateurs dans une zone encombrée. En particulier, pendant des périodes de congestion le transfert de données entre certains nœuds mobiles peut être difficile. En sachant que les utilisateurs vont rester dans la proximité d'un nœud cible (par la prédiction de κ-contact) l'opérateur de télécommunications peut décider d'utiliser un autre nœud mobile comme un proxy dans la communication avec l'utilisateur mobile ciblé.

Appendix B

List of Publications

- *A survey on predicting the popularity of web content* by Alexandru Tatar, Marcelo Dias de Amorim, Serge Fdida, and Panayotis Antoniadis. Journal of Internet Services and Applications (JISA), Springer, 2014
- *From Popularity Prediction to Ranking Online News* by Alexandru Tatar, Panayotis Antoniadis, Marcelo Dias de Amorim, and Serge Fdida, Social Network Analysis and Mining, Springer, 2014
- *Beyond Contact Predictions in Mobile Opportunistic Networks* by Alexandru Tatar, Tiphaine Phe-Neau, Marcelo Dias de Amorim, Vania Conan, and Serge Fdida. IFIP/IEEE Annual Conference on Wireless On-demand Network Systems and Services (WONS), Obergurgl, Austria, 2014
- *Ranking news articles based on popularity prediction* by Alexandru Tatar, Panayotis Antoniadis, Marcelo Dias de Amorim, and Serge Fdida. IEEE/ACM International Conference on Social Networks Analysis and Mining (ASONAM), Istanbul, Turkey, August 2012, (short paper)
- *Predicting the popularity of online articles based on user comments* by Alexandru Tatar, Panayotis Antoniadis, Marcelo Dias de Amorim, Jérémie Leguay, Arnaud Limbourg, and Serge Fdida. International Workshop on Social Data Mining for Human Behaviour Analysis, Sogndal, Norway, May 2011

List of figures

1.1 People gathering in St. Peter's Square eight years apart (Source: NBC[1]). . 16

1.2 The global scenario considered throughout this work and composed of a content producer, located on the Internet, a telecom operator that provides the infrastructure for the communication between mobile users and the content provider, and a set of collocated mobile users. 18

2.1 Data sets used as case-studies to evaluate the performance of prediction methods. On a log-log scale we depict the total number of items and the cumulative time period covered by each data set (using a **w**eekly, **m**onthly, **y**early demarcation). 25

2.2 A classification of web content popularity prediction methods. 30

2.3 Example of three popularity evolution trends discovered by Crane and Sornette [35] (and similar with some of the trends presented by Figueiredo [36] and Gursun et al. [29]). The figure shows the average number of views centered and normalized by the popularity during the peak day. 36

3.1 The number of articles and comments posted per hourly cycles. On the y-axis we illustrate the average number of articles and comments published per hour. 52

3.2 Complementary cumulative distribution function corresponding to the articles' lifetime (time elapsed between article publication time and the last comment time). The labels on the x-axis correspond to one hour, day, week, month, and year. We represent two versions of `20minutes` data: one over the entire data set and a reduced version that covers the same period of time as `telegraaf` data set. 53

3.3 Probability distribution function of the comments time relative to the articles publication time. We represent the histogram covering a one day period along with the best probability fit, which in our case is best described by a log-normal distribution. 54

3.4 The complementary cumulative distribution function of the articles' popularity and the corresponding power-law fit. 55

3.5 Normalized article ranks and the cumulative of proportion of comments received on a daily basis. We present the average value and one standard deviation (shaded area). 57

3.6 The prediction error in terms of QSE for the two popularity prediction methods. On the x-axis we vary the observation period from 1 to 24 hours. On the y-axis we represent the mean error (depicted in the top figures) and the mean along with one standard deviation represented by the shaded area in the bottom figures. 58

3.7 The prediction error in terms of QSE for the two popularity prediction methods. On the x-axis we vary the observation period from 1 to 24 hours. On the y-axis we represent the mean error (depicted in the top figures) and the mean along with one standard deviation represented by the shaded area in the bottom figures. 59

3.8 NDCG at different levels of precision. @n corresponds to the NDCG score for the top most important n articles. We present the mean over all prediction hours h (n=24) along with a 95% confidence interval. 60

3.9 Ranking accuracy in terms of NDCG@100 per hourly basis. The outer numbers correspond to different reference hours h (only the even hours of the day). The inner numbers correspond to the different ranking methods, with 1 - linear log, 2 - weighted, 3 - constant scaling, 4 - recency, 5 - live. 63

4.1 The global scenario considered composed of a content producer, located on the Internet, a telecom operator that provides the infrastructure for the communication between mobile users and the content provider, and a group of collocated mobile users. 70
4.2 The probability distribution function for the request arrival times relative to the content publication time for MediSyn synthetic workload. We represent the histogram covering a one-week period. 73
4.3 The distribution of content request per hour, on a weekly basis, generated by MediSyn. 74
4.4 The performance of the different proactive seeding strategies using a request timeout of 60 seconds and for different values of the additional load. 76
4.5 The performance of the different proactive seeding strategies for an additional traffic of 10% and for different durations of the request timeout. . . . 77
4.6 For an additional credit of credit of 5%, the performance of the *Scheduler* strategy using a request timeout of 60 seconds, compared to the perfect prediction strategies under different values of request timeout. 78

5.1 *Sig09* example: current vision versus vicinity awareness. 83
5.2 *Sig09* end-to-end transmission opportunities. 84
5.3 Example of κ-vicinity. The 1-vicinity consists in in all nodes found at a 1-hop distance. The 2-vicinity consists in all i neighbor's whose shortest distance is less than 2 hops. 85
5.4 Pairwise minimum distance for *Infocom05*, *Sig09*, and *Rollernet*. 87
5.5 The proportion of time (relative to the duration of a mobility trace) that nodes spend at a certain distance from one another. To simplify the representation we delineate with a distinct color only the first five distances and represent the remaining distances under a unique color (yellow). The pairs are ordered by the proportion of time spent in κ-vicinity = 5. 88
5.6 Prediction performance for different time-window durations and by varying the number of training time-windows (past intervals). 93
5.7 The efficiency of predicting κ-contact relationships for different durations of the time-window. On the y-axis we represent the prediction performance and on the x-axis we vary the value of κ-contact from 1 to 7. 94

5.8 The percent of traffic with the infrastructure that can be reduced through κ-contact prediction and mobile opportunistic communications. On the y-axis we represent the traffic reduction compared to the case where content is sent to mobile users using only the infrastructure. On the x-axis we present different values for κ-contact. 96

6.1 Opportunistic mobile data offloading procedure. 103

A.1 Personnes rassemblées dans la place Saint-Pierre dans une différence de huit ans (Source: NBC[2]). 106

A.2 Le scénario global considéré tout au long de ce travail est composé d'un producteur de contenu, situé sur l'Internet, un opérateur de télécommunications qui fournit l'infrastructure pour la communication entre les utilisateurs mobiles et le fournisseur de contenu, et un ensemble d'utilisateurs mobiles qui se trouvent en proximité. 108

A.3 La procédure utilisée dans cette thèse pour le délestage opportuniste de trafic mobile de données. 119

List of tables

2.1 Summary of the popularity prediction methods. 41

3.1 Summary of the data sets analyzed in this paper. 51

3.2 Comparing the power-law fit against other alternative distributions. For each alternative distribution, we provide the *p-value* and the likelihood ratio test (LR). We consider a significance level of 0.1 for the *p-value* and display the significant values in bold. Positive values of the log-likelihood indicate that the power-law is a better fit model than the alternative distributions. . 56

3.3 Ranking accuracy in terms of NDCG for different levels of precision. We compare the *linear log* model and the learning to rank algorithms using different set of features: **b**asic and **e**nhanced. The bold value indicates the best performing algorithm for a specific precision level. 65

4.1 Summary of the content creation and users' requests. 75

4.2 Mobility datasets characteristics. 75

5.1 Data sets characteristics. 85

5.2 The average duration that nodes remain at a certain distance (in seconds). 86

5.3 The average duration of a κ-contact relationship (in seconds). 89

5.4 κ-contact number of intervals ($\times 1,000$). 90

5.5 Notation for the binary classification confusion matrix 92

References

[1] "Telegraph." http://www.telegraph.co.uk/technology/twitter/9945505/Twitter-in-numbers.html, 2013.

[2] "Facebook statistics." https://newsroom.fb.com/News, 2013.

[3] "Youtube statistics." http://www.youtube.com/yt/press/statistics.html, 2013.

[4] "Mckinsey industry report." http://www.mckinsey.com/insights/business echnology/big data the next frontier for innovation, 2011.

[5] A. Aijaz, H. Aghvami, and M. Amani, "A survey on mobile data offloading: technical and business perspectives," *Wireless Communications, IEEE*, vol. 20, no. 2, no. 2, pp. 104–112, 2013.

[6] A. Balasubramanian, R. Mahajan, and A. Venkataramani, "Augmenting mobile 3g using wifi," in *Proceedings of the 8th international conference on Mobile systems, applications, and services*, pp. 209–222, ACM, 2010.

[7] J. G. Andrews, H. Claussen, M. Dohler, S. Rangan, and M. C. Reed, "Femtocells: Past, present, and future," *Selected Areas in Communications, IEEE Journal on*, vol. 30, no. 3, no. 3, pp. 497–508, 2012.

[8] A. Chaintreau, P. Hui, J. Crowcroft, C. Diot, R. Gass, and J. Scott, "Impact of human mobility on opportunistic forwarding algorithms," *IEEE Transactions on Mobile Computing*, vol. 6, no. 6, no. 6, pp. 606–620, 2007.

[9] A. Clauset and N. Eagle, "Persistence and periodicity in a dynamic proximity network," *arXiv preprint arXiv:1211.7343*, 2012.

[10] S. Heimlicher and K. Salamatian, "Globs in the Primordial Soup – The Emergence of Connected Crowds in Mobile Wireless Networks," (Chicago, Illinois, USA), Sept. 2010.

[11] A.-K. Pietiläinen, E. Oliver, J. LeBrun, G. Varghese, and C. Diot, "Mobiclique: middleware for mobile social networking," in *Proceedings of the 2nd ACM workshop on Online social networks*, pp. 49–54, ACM, 2009.

[12] B. Han, P. Hui, V. A. Kumar, M. V. Marathe, J. Shao, and A. Srinivasan, "Mobile data offloading through opportunistic communications and social participation," *Mobile Computing, IEEE Transactions on*, vol. 11, no. 5, no. 5, pp. 821–834, 2012.

[13] J. Whitbeck, Y. Lopez, J. Leguay, V. Conan, and M. D. De Amorim, "Push-and-track: Saving infrastructure bandwidth through opportunistic forwarding," *Pervasive and Mobile Computing*, vol. 8, no. 5, no. 5, pp. 682–697, 2012.

[14] "Internet advertising bureau." http://www.iabuk.net/about/press/archive/uk-digital-adspend-up-125-to-almost-55bn, 2013.

[15] D. Agarwal, B.-C. Chen, and X. Wang, "Multi-faceted ranking of news articles using post-read actions," in *Proceedings of the 21st ACM international conference on Information and knowledge management*, pp. 694–703, ACM, 2012.

[16] Y. Lifshits, "Ediscope: Social analytics for online news," *Yahoo Labs*, 2010.

[17] F. Figueiredo, F. Benevenuto, and J. M. Almeida, "The tube over time: characterizing popularity growth of youtube videos," in *Proceedings of the fourth ACM international conference on Web search and data mining*, pp. 745–754, ACM, 2011.

[18] Y. Yao and A. Sun, "Are most-viewed news articles most-shared?,"

[19] C. Castillo, M. El-Haddad, J. Pfeffer, and M. Stempeck, "Characterizing the life cycle of online news stories using social media reactions," *arXiv preprint arXiv:1304.3010*, 2013.

[20] G. Chatzopoulou, C. Sheng, and M. Faloutsos, "A first step towards understanding popularity in youtube," in *INFOCOM IEEE Conference on Computer Communications Workshops*, pp. 1–6, IEEE, 2010.

[21] M. Cha, A. Mislove, B. Adams, and K. P. Gummadi, "Characterizing social cascades in flickr," in *Proceedings of the first workshop on Online social networks*, pp. 13–18, ACM, 2008.

[22] E. Sadikov, M. Medina, J. Leskovec, and H. Garcia-Molina, "Correcting for missing data in information cascades," in *Proceedings of the fourth ACM international conference on Web search and data mining*, pp. 55–64, ACM, 2011.

[23] J. Cheng, L. Adamic, P. A. Dow, J. M. Kleinberg, and J. Leskovec, "Can cascades be predicted?," in *Proceedings of the 23rd international conference on World wide web*, pp. 925–936, International World Wide Web Conferences Steering Committee, 2014.

[24] "Cisco visual networking index: Forecast and methodology." http://www.cisco.com/c/en/us/solutions/collateral/service-provider, 2014.

[25] H. Kwak, C. Lee, H. Park, and S. Moon, "What is twitter, a social network or a news media?," in *Proceedings of the 19th international conference on World wide web*, pp. 591–600, ACM, 2010.

[26] "Reynolds journalism institute: News consumption on mobile media." http://www.rjionline.org/research/rji-dpa-mobile-media-project/2013-q1-research-report-1, 2008.

[27] "Mashable." http://mashable.com/2011/12/31/youtube-in-2011, 2013.

[28] X. Cheng, C. Dale, and J. Liu, "Statistics and social network of youtube videos," in *Quality of Service, 2008. IWQoS 2008. 16th International Workshop on*, pp. 229–238, IEEE, 2008.

[29] G. Gursun, M. Crovella, and I. Matta, "Describing and forecasting video access patterns," in *INFOCOM, 2011 Proceedings IEEE*, pp. 16–20, IEEE, 2011.

[30] P. Gill, M. Arlitt, Z. Li, and A. Mahanti, "Youtube traffic characterization: a view from the edge," in *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, pp. 15–28, ACM, 2007.

[31] M. Zink, K. Suh, Y. Gu, and J. Kurose, "Watch global, cache local: Youtube network traffic at a campus network: measurements and implications," in *Electronic Imaging 2008*, pp. 681805–681805, International Society for Optics and Photonics, 2008.

[32] M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon, "Analyzing the video popularity characteristics of large-scale user generated content systems," *IEEE/ACM Transactions on Networking (TON)*, vol. 17, no. 5, no. 5, pp. 1357–1370, 2009.

[33] X. Cheng, C. Dale, and J. Liu, "Understanding the characteristics of internet short video sharing: Youtube as a case study," *arXiv preprint arXiv:0707.3670*, 2007.

[34] Z. Avramova, S. Wittevrongel, H. Bruneel, and D. De Vleeschauwer, "Analysis and modeling of video popularity evolution in various online video content systems: power-law versus exponential decay," in *Evolving Internet, 2009. INTERNET'09. First International Conference on*, pp. 95–100, IEEE, 2009.

[35] R. Crane and D. Sornette, "Robust dynamic classes revealed by measuring the response function of a social system," *Proceedings of the National Academy of Sciences*, vol. 105, no. 41, no. 41, pp. 15649–15653, 2008.

[36] F. Figueiredo, "On the prediction of popularity of trends and hits for user generated videos," in *Proceedings of the sixth ACM international conference on Web search and data mining*, pp. 741–746, ACM, 2013.

[37] L. Carlinet, T. Huynh, B. Kauffmann, F. Mathieu, L. Noirie, and S. Tixeuil, "Four months in daily motion: Dissecting user video requests," in *Wireless Communications and Mobile Computing Conference (IWCMC), 2012 8th International*, pp. 613–618, IEEE, 2012.

[38] M. Ahmed, S. Spagna, F. Huici, and S. Niccolini, "A peek into the future: predicting the evolution of popularity in user generated content," in *Proceedings of the sixth ACM international conference on Web search and data mining*, pp. 607–616, ACM, 2013.

[39] Z. Dezsö, E. Almaas, A. Lukács, B. Rácz, I. Szakadát, and A.-L. Barabási, "Dynamics of information access on the web," *Physical Review E*, vol. 73, no. 6, no. 6, p. 066132, 2006.

[40] G. Mishne and N. Glance, "Leave a reply: An analysis of weblog comments," in *Third annual workshop on the Weblogging ecosystem*, 2006.

[41] A. Tatar, P. Antoniadis, M. D. de Amorim, and S. Fdida, "Ranking news articles based on popularity prediction," in *Proceedings of the 2012 International Conference*

on Advances in Social Networks Analysis and Mining (ASONAM 2012), pp. 106–110, IEEE Computer Society, 2012.

[42] M. Tsagkias, W. Weerkamp, and M. De Rijke, "News comments: Exploring, modeling, and online prediction," in *Advances in Information Retrieval*, pp. 191–203, Springer, 2010.

[43] G. Szabo and B. A. Huberman, "Predicting the popularity of online content," *Communications of the ACM*, vol. 53, no. 8, no. 8, pp. 80–88, 2010.

[44] K. Lerman and R. Ghosh, "Information contagion: An empirical study of the spread of news on digg and twitter social networks.," *ICWSM*, vol. 10, pp. 90–97, 2010.

[45] C. Wang, M. Ye, and B. A. Huberman, "From user comments to on-line conversations," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 244–252, ACM, 2012.

[46] A. Kaltenbrunner, V. Gomez, and V. Lopez, "Description and prediction of slashdot activity," in *Web Conference, 2007. LA-WEB 2007. Latin American*, pp. 57–66, IEEE, 2007.

[47] C. Wallenta, M. Ahmed, I. Brown, S. Hailes, and F. Huici, "Analysing and modelling traffic of systems with highly dynamic user generated content," *University College London, Tech. Rep. RN/08/10*, 2008.

[48] V. Gómez, A. Kaltenbrunner, and V. López, "Statistical analysis of the social network and discussion threads in slashdot," in *Proceedings of the 17th international conference on World Wide Web*, pp. 645–654, ACM, 2008.

[49] S. Jamali and H. Rangwala, "Digging digg: Comment mining, popularity prediction, and social network analysis," in *Web Information Systems and Mining, 2009. WISM 2009. International Conference on*, pp. 32–38, IEEE, 2009.

[50] K. Lerman and A. Galstyan, "Analysis of social voting patterns on digg," in *Proceedings of the first workshop on Online social networks*, pp. 7–12, ACM, 2008.

[51] S. Tang, N. Blenn, C. Doerr, and P. Van Mieghem, "Digging in the digg social news website," *Multimedia, IEEE Transactions on*, vol. 13, no. 5, no. 5, pp. 1163–1175, 2011.

[52] P. Van Mieghem, N. Blenn, and C. Doerr, "Lognormal distribution in the digg online social network," *The European Physical Journal B*, vol. 83, no. 2, no. 2, pp. 251–261, 2011.

[53] L. Hong, O. Dan, and B. D. Davison, "Predicting popular messages in twitter," in *Proceedings of the 20th international conference companion on World wide web*, pp. 57–58, ACM, 2011.

[54] H. Ma, W. Qian, F. Xia, X. He, J. Xu, and A. Zhou, "Towards modeling popularity of microblogs," *Frontiers of Computer Science*, vol. 7, no. 2, no. 2, pp. 171–184, 2013.

[55] S. KONG, F. YE, and L. FENG, "Predicting future retweet counts in a microblog," *Journal of Computational Information Systems*, vol. 10, no. 4, no. 4, pp. 1393–1404, 2014.

[56] T. Zaman, E. B. Fox, and E. T. Bradlow, "A bayesian approach for predicting the popularity of tweets," *arXiv preprint arXiv:1304.6777*, 2013.

[57] J. G. Lee, S. Moon, and K. Salamatian, "Modeling and predicting the popularity of online contents with cox proportional hazard regression model," *Neurocomputing*, vol. 76, no. 1, no. 1, pp. 134–145, 2012.

[58] A. Oghina, M. Breuss, M. Tsagkias, and M. de Rijke, "Predicting imdb movie ratings using social media," in *Advances in information retrieval*, pp. 503–507, Springer, 2012.

[59] D. A. Shamma, J. Yew, L. Kennedy, and E. F. Churchill, "Viral actions: Predicting video view counts using synchronous sharing behaviors.," in *ICWSM*, 2011.

[60] P. Yin, P. Luo, M. Wang, and W.-C. Lee, "A straw shows which way the wind blows: ranking potentially popular items from early votes," in *Proceedings of the fifth ACM international conference on Web search and data mining*, pp. 623–632, ACM, 2012.

[61] I. H. Witten and E. Frank, *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2005.

[62] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to information retrieval*, vol. 1. Cambridge University Press Cambridge, 2008.

[63] T. Broxton, Y. Interian, J. Vaver, and M. Wattenhofer, "Catching a viral video," in *Data Mining Workshops (ICDMW), 2010 IEEE International Conference on*, pp. 296–304, IEEE, 2010.

[64] M. Tsagkias, W. Weerkamp, and M. De Rijke, "Predicting the volume of comments on online news stories," in *Proceedings of the 18th ACM conference on Information and knowledge management*, pp. 1765–1768, ACM, 2009.

[65] R. Bandari, S. Asur, and B. A. Huberman, "The pulse of news in social media: Forecasting popularity.," in *ICWSM*, 2012.

[66] A. Tatar, P. Antoniadis, M. D. de Amorim, and S. Fdida, "From popularity prediction to ranking online news," *Social Network Analysis and Mining*, Ianuary 2014.

[67] A. Tatar, J. Leguay, P. Antoniadis, A. Limbourg, M. D. de Amorim, and S. Fdida, "Predicting the popularity of online articles based on user comments," in *Proceedings of the International Conference on Web Intelligence, Mining and Semantics*, p. 67, ACM, 2011.

[68] S.-D. Kim, S.-H. Kim, and H.-G. Cho, "Predicting the virtual temperature of web-blog articles as a measurement tool for online popularity," in *Computer and Information Technology (CIT), 2011 IEEE 11th International Conference on*, pp. 449–454, IEEE, 2011.

[69] H. Pinto, J. M. Almeida, and M. A. Gonçalves, "Using early view patterns to predict the popularity of youtube videos," in *Proceedings of the sixth ACM international conference on Web search and data mining*, pp. 365–374, ACM, 2013.

[70] W. Maass, T. Natschläger, and H. Markram, "Real-time computing without stable states: A new framework for neural computation based on perturbations," *Neural computation*, vol. 14, no. 11, no. 11, pp. 2531–2560, 2002.

[71] T. Wu, M. Timmers, D. D. Vleeschauwer, and W. V. Leekwijck, "On the use of reservoir computing in popularity prediction," in *Evolving Internet (INTERNET), 2010 Second International Conference on*, pp. 19–24, IEEE, 2010.

[72] J. Yang and J. Leskovec, "Patterns of temporal variation in online media," in *Proceedings of the fourth ACM international conference on Web search and data mining*, pp. 177–186, ACM, 2011.

[73] T. Hastie, R. Tibshirani, J. Friedman, and J. Franklin, *The elements of statistical learning: data mining, inference and prediction*. Springer, 2 ed., 2009.

[74] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *science*, vol. 315, no. 5814, no. 5814, pp. 972–976, 2007.

[75] K. Lerman and T. Hogg, "Using a model of social dynamics to predict popularity of news," in *Proceedings of the 19th international conference on World wide web*, pp. 621–630, ACM, 2010.

[76] S. D. Roy, T. Mei, W. Zeng, S. Li, and I. Fellow, "Towards cross-domain learning for social video popularity prediction," *IEEE Transactions on Multimedia*, 2013.

[77] S. D. Roy, T. Mei, W. Zeng, and S. Li, "Socialtransfer: cross-domain transfer learning from social streams for media applications," in *Proceedings of the 20th ACM international conference on Multimedia*, pp. 649–658, ACM, 2012.

[78] L. Marujo, M. Bugalho, J. P. d. S. Neto, A. Gershman, and J. Carbonell, "Hourly traffic prediction of news stories," *arXiv preprint arXiv:1306.4608*, 2013.

[79] J. A. Berger and K. L. Milkman, "What makes online content viral?," *Available at SSRN 1528077*, 2009.

[80] M. J. Salganik, P. S. Dodds, and D. J. Watts, "Experimental study of inequality and unpredictability in an artificial cultural market," *science*, vol. 311, no. 5762, no. 5762, pp. 854–856, 2006.

[81] T. Hogg and G. Szabo, "Diversity of user activity and content quality in online communities.," in *ICWSM*, 2009.

[82] A. Brodersen, S. Scellato, and M. Wattenhofer, "Youtube around the world: geographic popularity of videos," in *Proceedings of the 21st international conference on World Wide Web*, pp. 241–250, ACM, 2012.

[83] K. Huguenin, A.-M. Kermarrec, K. Kloudas, and F. Taïani, "Content and geographical locality in user-generated content sharing systems," in *Proceedings of the 22nd international workshop on Network and Operating System Support for Digital Audio and Video*, pp. 77–82, ACM, 2012.

[84] J. Ratkiewicz, A. Flammini, and F. Menczer, "Traffic in social media i: paths through information networks," in *Social Computing (SocialCom), 2010 IEEE Second International Conference on*, pp. 452–458, IEEE, 2010.

[85] Y. Borghol, S. Ardon, N. Carlsson, D. Eager, and A. Mahanti, "The untold story of the clones: content-agnostic factors that impact youtube video popularity," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1186–1194, ACM, 2012.

[86] J. Davidson, B. Liebald, J. Liu, P. Nandy, T. Van Vleet, U. Gargi, S. Gupta, Y. He, M. Lambert, B. Livingston, *et al.*, "The youtube video recommendation system," in *Proceedings of the fourth ACM conference on Recommender systems*, pp. 293–296, ACM, 2010.

[87] R. Zhou, S. Khemmarat, and L. Gao, "The impact of youtube recommendation system on video views," in *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*, pp. 404–410, ACM, 2010.

[88] R. Zhou, S. Khemmarat, L. Gao, and H. Wang, "Boosting video popularity through recommendation systems," in *Databases and Social Networks*, pp. 13–18, ACM, 2011.

[89] E. P. Markatos and C. E. Chronaki, "A top-10 approach to prefetching on the web," in *Proceedings of INET*, vol. 98, pp. 276–290, 1998.

[90] A. Kaltenbrunner, V. Gómez, A. Moghnieh, R. Meza, J. Blat, and V. López, "Homogeneous temporal activity patterns in a large online communication space," *CoRR*, 2007.

[91] G. Szabo and B. A. Huberman, "Predicting the popularity of online content," *Communications of the ACM*, vol. 53, no. 8, no. 8, p. 80, 2008.

[92] M. Tsagkias, W. Weerkamp, and M. De Rijke, "News comments: Exploring, modeling, and online prediction," in *Proceedings of the 32nd European conference on Advances in Information Retrieval*, ECIR2010, Springer, 2010.

[93] K. Lerman and T. Hogg, "Using a model of social dynamics to predict popularity of news," in *Proceedings of the 19th international conference on World wide web*, WWW '10, (New York, NY, USA), pp. 621–630, ACM, 2010.

[94] J. Lee, S. Moon, and K. Salamatian, "An approach to model and predict the popularity of online contents with explanatory factors," in *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01*, IEEE Computer Society, 2010.

[95] G. De Francisci Morales, A. Gionis, and C. Lucchese, "From chatter to headlines: harnessing the real-time web for personalized news recommendation," in *Proceedings of the fifth ACM international conference on Web search and data mining*, pp. 153–162, ACM, 2012.

[96] R. McCreadie, C. Macdonald, and I. Ounis, "News article ranking: Leveraging the wisdom of bloggers," in *Adaptivity, Personalization and Fusion of Heterogeneous Information*, pp. 40–48, 2010.

[97] A. Tatar, P. Antoniadis, A. Limbourg, M. D. de Amorim, J. Leguay, and S. Fdida, "Predicting the popularity of online articles based on user comments," in *WIMS'11*, pp. 67–75, ACM, 2011.

[98] M. Tsagkias, W. Weerkamp, and M. De Rijke, "Predicting the volume of comments on online news stories," in *Proceeding of the 18th ACM conference on Information and knowledge management*, pp. 1765–1768, ACM, 2009.

[99] M. Cha, H. Kwak, P. Rodriguez, Y. Ahn, and S. Moon, "I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system," in *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, pp. 1–14, ACM, 2007.

[100] M. Cha, A. Mislove, and K. P. Gummadi, "A measurement-driven analysis of information propagation in the flickr social network," in *Proceedings of the 18th international conference on World wide web*, pp. 721–730, ACM, 2009.

[101] M. Simkin and V. Roychowdhury, "Why does attention to web articles fall with time?," *Arxiv preprint arXiv:1202.3492*, 2012.

[102] F. Wu and B. Huberman, "Novelty and collective attention," *Proceedings of the National Academy of Sciences*, vol. 104, no. 45, no. 45, p. 17599, 2007.

[103] J. Leskovec, L. Backstrom, and J. Kleinberg, "Meme-tracking and the dynamics of the news cycle," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '09, ACM, 2009.

[104] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker, "Web caching and zipf-like distributions: Evidence and implications," in *INFOCOM'99. Eighteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE*, vol. 1, pp. 126–134, IEEE, 1999.

[105] L. Guo, E. Tan, S. Chen, Z. Xiao, and X. Zhang, "The stretched exponential distribution of internet media access patterns," in *Proceedings of the twenty-seventh ACM symposium on Principles of distributed computing*, pp. 283–294, ACM, 2008.

[106] S. Fortunato, A. Flammini, F. Menczer, and A. Vespignani, "Topical interests and the mitigation of search engine bias," *Proceedings of the National Academy of Sciences*, vol. 103, no. 34, no. 34, pp. 12684–12689, 2006.

[107] A. Clauset, C. Shalizi, and M. Newman, "Power-law distributions in empirical data," *SIAM Rev.*, vol. 51, pp. 661–703, November 2009.

[108] M. Mitzenmacher, "A brief history of generative models for power law and lognormal distributions," *Internet mathematics*, vol. 1, no. 2, no. 2, pp. 226–251, 2004.

[109] D. Easley and J. Kleinberg, *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge University Press, 2010.

[110] T.-Y. Liu, "Learning to rank for information retrieval," *Foundations and Trends in Information Retrieval*, vol. 3, no. 3, no. 3, pp. 225–331, 2009.

[111] K. Järvelin and J. Kekäläinen, "Cumulated gain-based evaluation of ir techniques," *ACM Transactions on Information Systems (TOIS)*, vol. 20, no. 4, no. 4, pp. 422–446, 2002.

[112] J. H. Friedman, "Greedy function approximation: a gradient boosting machine.(english summary)," *Ann. Statist*, vol. 29, no. 5, no. 5, pp. 1189–1232, 2001.

[113] Y. Freund, R. Iyer, R. E. Schapire, and Y. Singer, "An efficient boosting algorithm for combining preferences," *The Journal of Machine Learning Research*, vol. 4, pp. 933–969, 2003.

[114] Q. Wu, C. J. Burges, K. M. Svore, and J. Gao, "Adapting boosting for information retrieval measures," *Information Retrieval*, vol. 13, no. 3, no. 3, pp. 254–270, 2010.

[115] J. Xu and H. Li, "Adarank: a boosting algorithm for information retrieval," in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 391–398, ACM, 2007.

[116] V. Dang, "Ranklib library." http://people.cs.umass.edu/vdang/ranklib.html, 2012.

[117] A. J. Mashhadi and P. Hui, "Proactive caching for hybrid urban mobile networks," *University College London, Tech. Rep*, 2010.

[118] F. Malandrino, M. Kurant, A. Markopoulou, C. Westphal, and U. C. Kozat, "Proactive seeding for information cascades in cellular networks," in *INFOCOM, 2012 Proceedings IEEE*, pp. 1719–1727, IEEE, 2012.

[119] I. Trestian, S. Ranjan, A. Kuzmanovic, and A. Nucci, "Taming user-generated content in mobile networks via drop zones," in *INFOCOM, 2011 Proceedings IEEE*, pp. 2840–2848, IEEE, 2011.

[120] K. Lee, J. Lee, Y. Yi, I. Rhee, and S. Chong, "Mobile data offloading: how much can wifi deliver?," in *Proceedings of the 6th International COnference*, p. 26, ACM, 2010.

[121] R. Bhatia, G. Narlikar, I. Rimac, and A. Beck, "Unap: user-centric network-aware push for mobile content delivery," in *INFOCOM 2009, IEEE*, pp. 2034–2042, IEEE, 2009.

[122] K. Thilakarathna, A. C. Viana, A. Seneviratne, H. Petander, *et al.*, "The power of hood friendship for opportunistic content dissemination in mobile social networks," 2012.

[123] M. V. Barbera, J. Stefa, A. C. Viana, M. D. De Amorim, and M. Boc, "Vip delegation: Enabling vips to offload data in wireless social mobile networks," in *Distributed Computing in Sensor Systems and Workshops (DCOSS), 2011 International Conference on*, pp. 1–8, IEEE, 2011.

[124] F. Rebecchi, M. Dias de Amorim, and V. Conan, "DROiD: adapting to individual mobility pays off in mobile data offloading," in *IFIP Networking Conference (Networking 2014)*, 2014.

[125] P. Baier, F. Durr, and K. Rothermel, "Tomp: Opportunistic traffic offloading using movement predictions," in *Local Computer Networks (LCN), 2012 IEEE 37th Conference on*, pp. 50–58, IEEE, 2012.

[126] A. Noori and D. Giustiniano, "Hycloud: a hybrid approach toward offloading cellular content through opportunistic communication," in *Proceeding of the 11th annual international conference on Mobile systems, applications, and services*, pp. 551–552, ACM, 2013.

[127] Y. Li, G. Su, P. Hui, D. Jin, L. Su, and L. Zeng, "Multiple mobile data offloading through delay tolerant networks," in *Proceedings of the 6th ACM workshop on Challenged networks*, pp. 43–48, ACM, 2011.

[128] T. Wang, P. Hui, S. R. Kulkarni, and P. Cuff, "Cooperative caching based on file popularity ranking in delay tolerant networks," *Proc. Of ExtremeCom*, 2012.

[129] J. Famaey, F. Iterbeke, T. Wauters, and F. DeTurck, "Towards a predictive cache replacement strategy for multimedia content," *Journal of Network and Computer Applications*, 2012.

[130] U. Paul, A. P. Subramanian, M. M. Buddhikot, and S. R. Das, "Understanding traffic dynamics in cellular data networks," in *INFOCOM, 2011 Proceedings IEEE*, pp. 882–890, IEEE, 2011.

[131] M. Z. Shafiq, L. Ji, A. X. Liu, and J. Wang, "Characterizing and modeling internet traffic dynamics of cellular devices," in *Proceedings of the ACM SIGMETRICS joint international conference on Measurement and modeling of computer systems*, pp. 305–316, ACM, 2011.

[132] W. Tang, Y. Fu, L. Cherkasova, and A. Vahdat, "Medisyn: A synthetic streaming media service workload generator," in *Proceedings of the 13th international workshop on Network and operating systems support for digital audio and video*, pp. 12–21, ACM, 2003.

[133] P.-U. Tournoux, J. Leguay, F. Benbadis, J. Whitbeck, V. Conan, and M. D. de Amorim, "Density-aware routing in highly dynamic DTNs: The rollernet case," *IEEE Transactions on Mobile Computing*, vol. 10, pp. 1755–1768, 2011.

[134] M. Salathé, M. Kazandjieva, J. W. Lee, P. Levis, M. W. Feldman, and J. H. Jones, "A high-resolution human contact network for infectious disease transmission," *Proceedings of the National Academy of Sciences*, vol. 107, no. 51, no. 51, pp. 22020–22025, 2010.

[135] E. Cohen and S. Shenker, "Replication strategies in unstructured peer-to-peer networks," in *ACM SIGCOMM Computer Communication Review*, vol. 32, pp. 177–190, ACM, 2002.

[136] V. Conan, J. Leguay, and T. Friedman, "Characterizing pairwise inter-contact patterns in delay tolerant networks," in *Proceedings of the 1st international conference on Autonomic computing and communication systems*, p. 19, ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2007.

[137] A. Passarella and M. Conti, "Characterising aggregate inter-contact times in heterogeneous opportunistic networks," in *NETWORKING 2011*, pp. 301–313, Springer, 2011.

[138] E. Borgia, M. Conti, and A. Passarella, "Autonomic detection of dynamic social communities in opportunistic networks," in *Ad Hoc Networking Workshop (Med-Hoc-Net), 2011 The 10th IFIP Annual Mediterranean*, pp. 142–149, IEEE, 2011.

[139] M. Mordacchini, A. Passarella, and M. Conti, "Community detection in opportunistic networks using memory-based cognitive heuristics," in *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2014 IEEE International Conference on*, pp. 243–248, IEEE, 2014.

[140] C. Boldrini, M. Conti, and A. Passarella, "Exploiting users social relations to forward data in opportunistic networks: The hibop solution," *Pervasive and Mobile Computing*, vol. 4, no. 5, no. 5, pp. 633–657, 2008.

[141] V. Erramilli, A. Chaintreau, M. Crovella, and C. Diot, "Diversity of forwarding paths in pocket switched networks," in *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, pp. 161–174, ACM, 2007.

[142] V. Erramilli, M. Crovella, A. Chaintreau, and C. Diot, "Delegation forwarding," in *Proceedings of the 9th ACM international symposium on Mobile ad hoc networking and computing*, pp. 251–260, ACM, 2008.

[143] M.-H. Zayani, V. Gauthier, and D. Zeghlache, "Improving link prediction in intermittently connected wireless networks by considering link and proximity stabilities," in *WOWMOM*, pp. 1–10, 2012.

[144] T. Phe-Neau, *Properties and impact of vicinity in mobile opportunistic networks.* PhD thesis, Université Pierre et Marie Curie, 2014.

[145] T. Phe-Neau, M. Dias de Amorim, and V. Conan, "The Strength of Vicinity Annexation in Opportunistic Networking," (Torino, Italy), Apr. 2013.

[146] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási, "Limits of predictability in human mobility," *Science*, vol. 327, no. 5968, no. 5968, pp. 1018–1021, 2010.

[147] D. Liben-Nowell and J. Kleinberg, "The link-prediction problem for social networks," *J. Am. Soc. Inf. Sci. Technol.*, vol. 58, no. 7, pp. 1019–1031, May 2007.

[148] D. Wang, D. Pedreschi, C. Song, F. Giannotti, and A.-L. Barabasi, "Human mobility, social ties, and link prediction," pp. 1100–1108, 2011.

[149] B. Taskar, M.-F. Wong, P. Abbeel, and D. Koller, "Link prediction in relational data," in *Neural Information Processing Systems*, vol. 15, 2003.

[150] Z. Huang and D. K. Lin, "The time-series link prediction problem with applications in communication surveillance," *INFORMS Journal on Computing*, vol. 21, no. 2, no. 2, pp. 286–303, 2009.

[151] M. Al Hasan, V. Chaoji, S. Salem, and M. Zaki, "Link prediction using supervised learning," in *SDM 06: Workshop on Link Analysis, Counter-terrorism and Security*, 2006.

[152] T. Phe-Neau, M. Dias de Amorim, and V. Conan, "Vicinity-based DTN Characterization," (Zurich, Switzerland), Mar. 2012.

[153] A.-K. Pietiläinen and C. Diot, "Dissemination in opportunistic social networks: the role of temporal communities," (Hilton Head, South Carolina, USA), June 2012.

[154] Z. Huang, X. Li, and H. Chen, "Link prediction approach to collaborative filtering," in *Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries*, pp. 141–142, ACM, 2005.

[155] L. A. Adamic and E. Adar, "Friends and Neighbors on the Web," *Social Networks*, vol. 25, no. 3, no. 3, pp. 211–230, 2003.

[156] L. Katz, "A new status index derived from sociometric analysis," *Psychometrika*, vol. 18, no. 1, no. 1, pp. 39– 43, 1953.

[157] A.-L. Barabâsi, H. Jeong, Z. Néda, E. Ravasz, A. Schubert, and T. Vicsek, "Evolution of the social network of scientific collaborations," *Physica A: Statistical Mechanics and its Applications*, vol. 311, no. 3, no. 3, pp. 590–614, 2002.

[158] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[159] A. Keränen, J. Ott, and T. Kärkkäinen, "The ONE Simulator for DTN Protocol Evaluation," (Rome, Italy), Mar. 2009.

[160] "Ko-tag joint project." http://ko-fas.de/english/ko-tag—cooperative-transponders/operating-principle.html/, 2014.

[161] J. Leskovec, L. Backstrom, and J. Kleinberg, "Meme-tracking and the dynamics of the news cycle," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 497–506, ACM, 2009.

[162] S. Nikolov, *Trend or no trend: a novel nonparametric method for classifying time series*. PhD thesis, Massachusetts Institute of Technology, 2012.

[163] K. Radinsky, S. Davidovich, and S. Markovitch, "Predicting the news of tomorrow using patterns in web search queries," in *Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology-Volume 01*, pp. 363–367, IEEE Computer Society, 2008.

[164] K. Radinsky, S. Davidovich, and S. Markovitch, "Learning causality for news events prediction," in *Proceedings of the 21st international conference on World Wide Web*, pp. 909–918, ACM, 2012.

[165] K. Radinsky, K. Svore, S. Dumais, J. Teevan, A. Bocharov, and E. Horvitz, "Modeling and predicting behavioral dynamics on the web," in *Proceedings of the 21st international conference on World Wide Web*, pp. 599–608, ACM, 2012.

[166] T. Steiner, S. van Hooland, and E. Summers, "Mj no more: using concurrent wikipedia edit spikes with social network plausibility checks for breaking news detection," in *Proceedings of the 22nd international conference on World Wide Web companion*, pp. 791–794, International World Wide Web Conferences Steering Committee, 2013.

[167] T. Steiner, "Telling breaking news stories from wikipedia with social multimedia: A case study of the 2014 winter olympics," *arXiv preprint arXiv:1403.4289*, 2014.

[168] "Incoming tv media application." http://http://www.incoming-media.com/, 2014.

[169] J. Whitbeck, M. Dias de Amorim, V. Conan, and J.-L. Guillaume, "Temporal reachability graphs," in *Proceedings of the 18th annual international conference on Mobile computing and networking*, pp. 377–388, ACM, 2012.