# INDIGO: A Generalized Model and Framework for Performance Prediction of Data Dissemination

Doctoral Dissertation submitted to the
Faculty of Informatics of the Università della Svizzera Italiana
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

presented by
Kamini Garg

under the supervision of
Prof. Mehdi Jazayeri and Prof. Silvia Giordano

February 2017

Dissertation Committee

| | |
|---|---|
| **Prof. Sajal K. Das** | Missouri University of Science and Technology, USA |
| **Prof. Oscar Mayora** | Fondazione Bruno Kessler–CREATE-NET, Trento, Italy |
| **Prof. Fabio Crestani** | Università della Svizzera Italiana, Switzerland |
| **Prof. Cesare Pautasso** | Università della Svizzera Italiana, Switzerland |

Dissertation accepted on 27 February 2017

| Research Advisor | Co-Advisor |
|---|---|
| **Prof. Mehdi Jazayeri** | **Prof. Silvia Giordano** |

PhD Program Director

**Prof. Walter Binder, Prof. Michael Bronstein**

I certify that except where due acknowledgement has been given, the work presented in this thesis is that of the author alone; the work has not been submitted previously, in whole or in part, to qualify for any other academic award; and the content of the thesis is the result of work which has been carried out since the official commencement date of the approved research program.

Kamini Garg
Lugano, 27 February 2017

*To Vidya Buaji*

# Abstract

According to recent studies, an enormous rise in location-based mobile services is expected in future. People are interested in getting and acting on the localized information retrieved from their vicinity like local events, shopping offers, local food, etc. These studies also suggested that local businesses intend to maximize the reach of their localized offers/advertisements by pushing them to the maximum number of interested people. The scope of such localized services can be augmented by leveraging the capabilities of smartphones through the dissemination of such information to other interested people.

To enable local businesses (or publishers) of localized services to take informed decision and assess the performance of their dissemination-based localized services in advance, we need to predict the performance of data dissemination in complex real-world scenarios. Some of the questions relevant to publishers could be the maximum time required to disseminate information, best relays to maximize information dissemination etc. This thesis addresses these questions and provides a solution called INDIGO that enables the prediction of data dissemination performance based on the availability of physical and social proximity information among people by collectively considering different real-world aspects of data dissemination process.

INDIGO empowers publishers to assess the performance of their localized dissemination based services in advance both in physical as well as the online social world. It provides a solution called *INDIGO–Physical* for the cases where physical proximity plays the fundamental role and enables the tighter prediction of data dissemination time and prediction of best relays under real-world mobility, communication and data dissemination strategy aspects. Further, this thesis also contributes in providing the performance prediction of data dissemination in large-scale online social networks where the social proximity is prominent using *INDIGO–OSN* part of the INDIGO framework under different real-world dissemination aspects like heterogeneous activity of users, type of information that needs to be disseminated, friendship ties and the content of the published online activities.

INDIGO is the first work that provides a set of solutions and enables publishers to predict the performance of their localized dissemination based services based on the availability of physical and social proximity information among people and different real-world aspects of data dissemination process in both physical and online social networks. INDIGO outperforms the existing works for physical proximity by providing 5 times tighter upper bound of data dissemination time under real-world data dissemination aspects. Further, for social proximity, INDIGO is able to predict the data dissemination with 90% accuracy and differently, from other works, it also provides the trade-off between high prediction accuracy and privacy by introducing the feature planes from an online social networks.

# Acknowledgements

I would like to first express my gratitude to my advisors Prof. Silvia Giordano and Prof. Mehdi Jazayeri. I would like to thank Prof. Silvia Giordano for providing me the opportunity to start my Ph.D. and introducing me into the research area of mobile and social computing. She has always supported and motivated me during my Ph.D. and always gave me her valuable feedback for advancing my research. I would like to thank Prof. Mehdi Jazayeri for his constructive comments and valuable feedback to my work that helped me to keep the focus on my research. A special thanks go to all the members of the dissertation committee for providing their constructive and useful comments during my research proposal.

Many thanks go to all the people I had the pleasure of collaborating with: Aleksandar Matic, Souneil Park and Nuria Oliver for providing me an opportunity to do my internship at Telefónica Research and collaborating with me for the social proximity part of my thesis. I also thank Katia JAFFRES-RUNSER and Aline Carneiro Viana for their collaboration of MACACO traces. Further, I thank Valerio Arnaboldi for collaborating the Online Social Networks part of my thesis and Prof. Anna Förster for her support during the initial phase of my Ph.D. Further, I would like to thank my colleagues at SUPSI for their support. A special thanks go to Michela Papandrea, Alan Ferrari, Salvatore Vanini and Daniele Puccinelli. I also thank Dario Gallucci for his collaboration during the PerCom data collection.

I am extremely grateful to my parents and to my brother and sisters, for their constant and unconditional love and support. Finally and most importantly, I give my deepest and most special thanks to my beloved husband Steven for his immense support and love during my Ph.D.

# Contents

# Figures

# Tables

# Chapter 1

# Introduction

The immense growth of smartphones has enabled to provide ubiquitous reachability to various types of information, enhanced by powerful communication and computing resources, and intuitive user experience. These mobile devices have quickly become a part of our everyday lives and the proliferation of these devices has not just altered the way we communicate and interact, but it has also led to a significant innovation of services. It has opened a door to a new set of applications such as location-based advertising, recommendations entertainment, health care etc. and has also made it possible to gather contextual and personalized information. Recent studies have highlighted an enormous rise in location-based search, advertisements, and services Google [2015]. People are usually interested in getting and taking advantage of localized information received from their vicinity like local events, shopping offers, local food, transport, traffic information etc. Further, local businesses also intend to maximize the reach of their localized offers/advertisements by pushing them to the maximum number of interested people. The scope of such localized services can be augmented by leveraging the capabilities of smartphones through the dissemination of such information to other interested people. The effectiveness of such techniques to enable fast and efficient information dissemination has also been suggested in literature Dimatteo et al. [2011] Whitbeck et al. [2011]. The impact of these localized information gathering and dissemination is two-fold:

- It can help both local businesses (or publishers) to further disseminate their information to other geographical regions.
- It can bring information close to the people who are not aware of it but might be interested in it.

Before deploying any new localized offers or advertisements in a given ge-

ographical region, some of the key questions that can arise from the publisher point of view could be:

- How relevant my localized offer is among people?
- How much time is required to disseminate information to a certain fraction of interested people?
- Who are the people who can maximize the information dissemination?

For example, if the maximum time to disseminate information to all or fraction of interested people tends to infinity, then the offered service will be useless due to its inability to spread information to people within a reasonable time limit. To answer such questions and enable publishers to assess the performance of their localized service, we need to predict the performance of data dissemination in such complex mobile networks that exhibit heterogeneous contact patterns and where the interest of people is a key feature. To predict the performance of data dissemination realistically and empower local publishers to take informed decisions about their services, there is a need for the unified model that enables the prediction of data dissemination under multiple consideration of different real-world aspects such as heterogeneous and time-varying contact patterns, interests of people, different data dissemination strategies, multiple-simultaneous contacts among people, data originating from multiple sources etc.. Existing works do not collectively consider these real-world aspects of data dissemination and fail to provide a solution to model data dissemination process for different performance metrics.

In this thesis, I propose a solution called INDIGO, a generalized data dissemination framework that enables the prediction of data dissemination by *collectively* considering the real-world aspects of data dissemination under the availability of both physical and social proximity information as presented in Physical-Social proximity table (see Figure 1.1). The physical proximity represents the closeness of people in the physical world while social proximity shows their closeness according to their interests or social network information. As illustrated in Figure 1.1, data dissemination can happen also without physical proximity information in case of online social networks. To tackle this aspect, I further present a solution to predict the performance of data dissemination in online social networks like Twitter where social proximity dominates over the physical proximity as the part of the INDIGO framework.

Altogether, my thesis proposed a solution called INDIGO that provides solutions in case of availability of physical and social proximity information. INDIGO

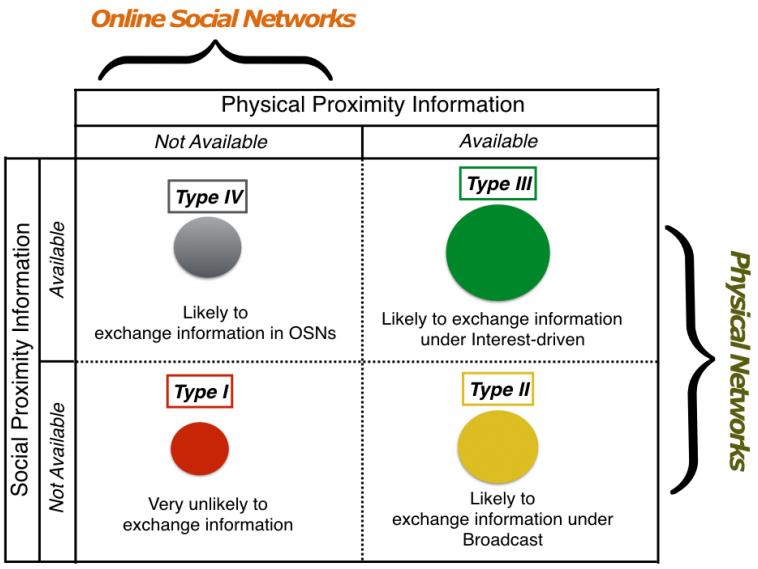



Figure 1.1. Modeling of data dissemination process based on the Physical-Social proximity information availability. In the presence of physical proximity information (Type II) or both physical and social proximity information (Type III), the data gets exchanged in physical networks under broadcast and interest-driven data dissemination strategy respectively. In case when only social proximity information is available (Type IV), the data is shared on online social networks because Type IV users are generally linked with a online social networks.

covers all aspects of data dissemination, from data exchanging, that happens when two users physically encounter each other, up to data sharing, which happens when an user disseminate data to another user via a online social link. Now-a-days, both dissemination types, data exchanging and data sharing, happen in a very interleaved way, and it is of paramount importance to consider both of them while predicting the performance of data dissemination.

## 1.1 Problem Statement

This thesis focuses on providing a solution to enable pre-deployment performance analysis of dissemination-based localized services by predicting the performance of data dissemination under real-world aspects of data dissemination process considering the availability of *physical* and *social* proximity information. The thesis will be mainly concerned with the following research questions:

- What real-world aspects should we consider to enable realistic performance prediction of dissemination-based localized services in case of the availability of physical and/or social proximity information?
- How to learn the real interests of people from their mobile devices?
- How to collectively consider and model different real-world mobility and social (like preferences and interests of people) aspects for performance prediction of data dissemination process under the availability of both physical and social proximity?
- Does INDIGO achieve realistic and tighter upper bounds of data dissemination process for different cases of physical and social proximity?
- What dissemination real-world aspects to consider and how to model data dissemination in online social networks where social proximity overpowers the physical proximity?

Therefore, this dissertation solves the problem of:

*"providing a solution to predict the performance of data dissemination by collectively considering the real-world aspects of data dissemination process based on the availability of physical and social proximity among people."*

## 1.2 Motivation: A solution to predict the performance of data dissemination

To provide realistic pre-deployment performance prediction of data dissemination process and empower publishers to assess the performance of their services, the proposed solution needs to collectively consider the real world-aspect of data dissemination process for different cases of physical and social proximity information avialability. In the case of physical networks where we can have either physical proximity or both physical and social proximity infomation, existing works have put a fundamental basis for the performance prediction of data dissemination process in mobile networks. The work done so far mainly look at the physical proximity case while considering mobility patterns of people Groenevelt et al. [2005] Mosk-Aoyama and Shah [2008] Pettarin et al. [2011] Peres et al. [2011] Picu et al. [2012] Passarella and Conti [2013]. In the case of both physical and social proximity, the real-world aspects are usually modeled by existing works either by looking at the friendship graph of people or they explicitly

ask users for their interests across certain topics Hui et al. [2008] Mei et al. [2011] Ciobanu et al. [2015].

The different approaches taken by all these works fail to provide the realistic prediction of data dissemination process and also differ for different data sets. Therefore, such solutions are neither suitable for publishers nor they provide a unique solution that collectively considers different key aspects of data dissemination process such as heterogeneous and time-varying contact patterns, multiple-simultaneous contacts among people, different data dissemination strategies & data requirements and modeling of data originating from different data sources Garg and Giordano [2015]. I highlight the importance of some key real-world aspects (mobility aspects) on data dissemination time and then motivate the need for a unified solution by showing their impact on data dissemination time using two benchmark traces used in literature i.e. INFOCOM and MIT contact traces (detailed description of these traces are presented in Chapter 3).

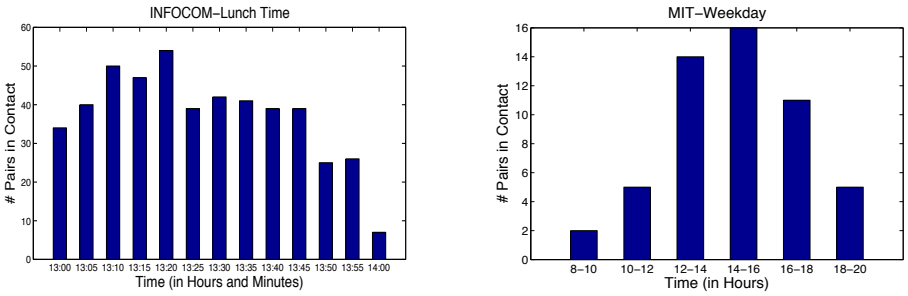
Figure 1.2 shows the occurrence of multiple simultaneous contacts in two diverse environments INFOCOM Scott et al. [2006c] (conference) and MIT Eagle et al. [2009] (university) at different timeslots. Further, I also calculate the total number of pairs with distinct contact probabilities in both trace and also inspects heterogeneity in the contact patterns of people through their dispersion from each other using Coefficient of Variation (CV). My analysis shows that both traces contain several distinct contact probabilities (INFOCOM: 687 and MIT: 749) with significant variability (INFOCOM: CV=1.25 and MIT: CV=1.11)[1] thus, highlights the presence of heterogeneous contacts.

***Observation 1:*** *Assumption of homogeneous contact probabilities among people is not realistic for the performance prediction of data dissemination process due to significant heterogeneity observed in the contact patterns of people in real world environments.*

***Observation 2:*** *The multiple simultaneous contacts occur between different pairs in different times, therefore, the assumption of sequential contacts while modeling data dissemination does not hold in reality.*

In addition to the heterogeneity in contact patterns, I also found the time-varying contact patterns shown in Figure 1.3. For INFOCOM trace, I calculate the total number of pair-wise contacts on an hourly basis and for MIT trace I calculate daily contacts among people for different weeks. In INFOCOM trace, I

[1] $CV > 1$ implies high variance and vice versa.




(a) INFOCOM (b) MIT

Figure 1.2. Number of simultaneous contact pairs for different time spans from INFOCOM and MIT traces.

observe a significant increase in the number of contacts during lunch time and coffee break. Similarly, from Figure 1.3(b), I observe that during the $3^{rd}$ month of MIT trace, the number of contacts during week days are significantly higher than those during weekends. It happens because in a university environment people are more likely to meet during weekdays (office hours) as compared to weekends.

***Observation 3:*** *Time and context play an important role in the contact patterns of people and will impact the performance prediction of data dissemination process.*

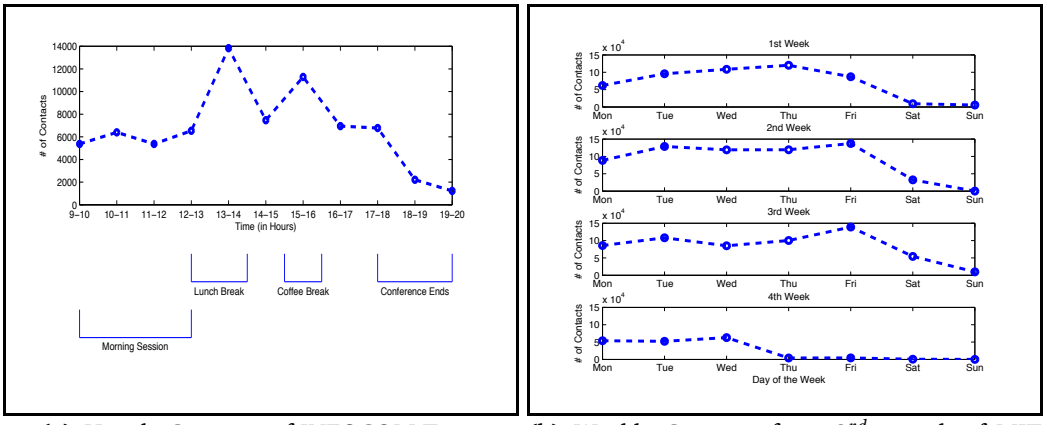
Finally, to show the impact of heterogeneous mobility and multiple simultaneous contacts on data dissemination process, I measure the data dissemination time under three cases:

1. $T_{Real}$*:* Real data dissemination time measured by utilizing contact traces that exhibits real-world mobility aspects (heterogeneous pair-wise contact probability and multiple simultaneous contacts among people). I call this data dissemination time as ground truth.

2. $T_{HOMO}$*:* Data dissemination time calculated by considering homogeneous contact probability (mean of all pair-wise contact probabilities) among all pairs of people.

3. $T_{NSC}$*:* Data dissemination time measured without considering multiple simultaneous contacts among people.

Figure 1.4 presents the data dissemination time obtained in all three cases for INFOCOM and MIT trace. Our analysis highlights the impact of multiple simultaneous contacts and pair-wise heterogeneous contact probabilities on data

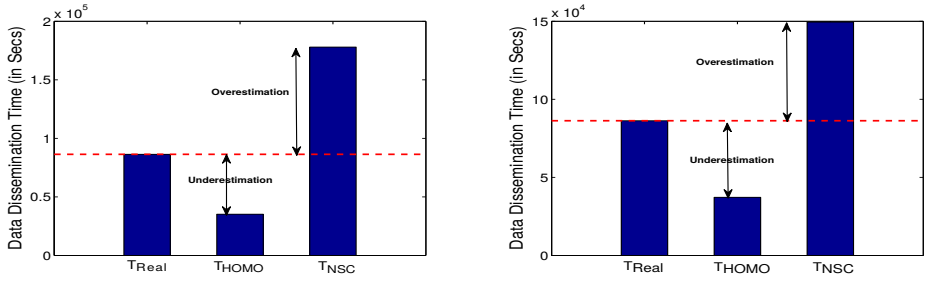



(a) Hourly Contacts of INFOCOM Trace (b) Weekly Contacts from $3^{rd}$ month of MIT Trace

Figure 1.3. Number of Hourly and Weekly contacts extracted from INFOCOM and MIT trace. The number of contacts show the impact of time and context on the contact patterns of people.

dissemination time. We observe that assumption of homogeneous contact probability underestimates data dissemination time ($T_{HOMO}$) while the assumption of sequential contacts among people overestimates data dissemination time ($T_{NSC}$). Our analysis shows that both assumptions are unable to mimic real data dissemination time ($T_{Real}$) thus, signifies the importance to collectively consider real-world mobility aspects.

***Observation 4:*** *Multiple simultaneous contacts occur among people in different environments and needs to be considered to improve the performance prediction of data dissemination process.*




(a) INFOCOM (b) MIT

Figure 1.4. Data dissemination time for all three cases ($T_{Real}$, $T_{HOMO}$ and $T_{NSC}$) in INFOCOM and MIT traces.

Above observations show the importance of collectively considering different aspects of data dissemination time while predicting their performance. Furthermore, as opposed to existing works, the proposed solution should not always rely on complete dataset rather it should be able to predict contact patterns and interests of people for future scenarios where data is unavailable. To enable publishers to take decision according to their requirement (or service requirement) and maximize the reach of their localized services, it is also of paramount importance to provide a solution that predicts data dissemination performance under difference cases of social and physical proximity considering the above described real-world aspects. In this thesis, I provide a solution that considers all these aspects of data dissemination process and help publishers to take informative decisions according to their requirement. Further, this thesis also takes into account the data sharing on online social networks by quantifying the different level of rich information available. The recent works have focused on utilizing different set of information like friendship graph, content informationGaluba et al. [2010]; Petrovic et al. [2011] Myers et al. [2012a] for the case of data dissemination in online social networks where we only have social proximity information availability.

## 1.3 The Goal: A generalized data dissemination framework

In this thesis, I offer a generalized framework called INDIGO to conduct predeployment performance analysis of data dissemination process in a given scenarios. INDIGO predicts the performance of data dissemination in multiple dimensions by collectively considering different real-world aspects of data dissemination process based on the availability of physical and social proximity. Further, INDIGO offers solutions to publishers to access the performance of their dissemination-based localized services based on the availability of physical and social proximity information. The contribution of this thesis is summarized as follows:

- I propose a framework INDIGO that enables the prediction of the performance of data dissemination in given scenario based on the availability of physical and social proximity information for all cases of Physical–Social proximity table (Type II, III and IV). To the best of my knowledge, INDIGO is the first work that provides solutions to different cases and predict the performance of data dissemination under different cases of proximity

by collectively consider different real-world aspects of data dissemination process.

- INDIGO helps publishers (or local businesses) to estimate the likelihood of success in their initiatives by providing a solution to conduct pre-deployment performance analysis of dissemination-based localized services/offers in physical as well as online social networks.

- INDIGO predicts the performance of data dissemination under real-world mobility characteristics like heterogeneous and time-varying contact patterns and multiple simultaneous contacts among people in case of the availability of the physical proximity information. Further, INDIGO also considers social information such as interests and preferences of people along with multiple data sources, different user data requirement and data dissemination strategies (like broadcast and interest-driven) for the physical networks case when both physical and social proximity information is available. Further, in case of social proximity information availability, the real-world dissemination aspects are the heterogeneous activity of users on online social networks, type of information that needs to be disseminated, friendship ties and the content of the published online activities.

- INDIGO also provide a machine learning based solution to predict the future pair-wise contact patterns of people and it also handles the automatic learning of real interests of people from the web browsing history of their Smartphones. In addition to this, it also provides a solution to generate artificial interests of people in case they are not available in the dataset.

- As opposed to existing works, the evaluation metrics obtained from INDIGO is not only limited to data dissemination time for physical networks. It can also find the best relays to maximize information spread or to minimize data dissemination time.

- Finally, INDIGO provides a complete solution to conduct pre-deployment analysis by learning and providing contact patterns, interest profiles under different data dissemination strategy, data requirements and multiple data sources for physical networks covering Type II and III.

- The results obtained from INDIGO for Type II and III for both upper bounds of data dissemination time and best relays are validated using benchmark traces. Further, I also conducted my own experiments to collect traces from

two diverse environments, a conference (PerCom 2012) and the university of two different countries (MACACO–France and (MACACO–Brazil) through a dedicated mobile application. The collected university traces captures both contact and real interests of people from their web browsing history. To the best of my knowledge, this is the first dataset that captures interests of people along with their real mobility patterns.

- Finally, INDIGO also provide a machine learning based model to predict the performance of data dissemination on online social networks where social proximity plays an important role. In this way, this thesis provide a solution to fill the gray area of Physical-Social proximity table of Figure 1.1. INDIGO also introduces multiple feature planes using the set of rich information available on Twitter dataset according to their complexity to acquire and privacy intrusiveness.

### 1.3.1 Use cases of INDIGO

INDIGO aims to provide support for the diverse applications. Some examples are described below:

#### Dissemination of local advertisements or social campaign

With the help of INDIGO, local businesses can find the upper bound of time required to send information to all or a fraction of people based on their interests. Also by finding the set of best relays, our framework can help publishers to boost the dissemination of their advertisements and campaigns.

#### Dissemination of traffic conditions and accidental information

INDIGO can also be useful for broadcasting the information about the traffic condition and accidental information in a given area. Further, it can also predict the best relays in the network such that this information would be disseminated to the maximum number of people in a limited amount of time. Thus, it can help commuters to take better decision while planning their journey.

#### Viral marketing on Online Social Networks

INDIGO can be useful to predict the success of a new product, campaign or advertisement of companies by predicting their dissemination on online social networks. The number of times a particular post or tweet is diffused in the network

is considered as the key performance indicator to measure the success of such events[2].

## 1.4 Contribution

The main contribution of this thesis provides a solution to predict the performance of data dissemination time for different cases of Physical-Social proximity table. The thesis covers all cases *Type II to Type IV* of the Physical-Social proximity table presented in Figure 1.1[3]. The physical proximity information is driven by the physical contacts/meeting of people while the social proximity information comes under the umbrella of willingness to share information based on their interests and preferences, friendship ties and their communication on online social networks. The explanation of different types of information exchange is presented as follows:

1. Type I: In this case, both physical and social proximity information is not available therefore, it is very unlikely to exchange information among people as we do not have any information about the people in any dimension.

2. Type II: In this case, the physical proximity information among people is available. In this situation, people are physically closer and can exchange information under broadcast strategy where the data dissemination is only driven by physical proximity information.

3. Type III: In this case, both physical and social proximity information is available in physical networks where social proximity information is considered using the interests of people. Therefore, in this case, the data dissemination will occur under interest-driven strategy where both contacts and interests of people are modeled for data dissemination.

4. Type IV: In this case, only the social proximity information is available among people. This case arises when people are connected with each other through online social networks where physical proximity is not the necessary requirement. Therefore, in this case, the information can only be shared online via social link using online social networks. Type IV case

[2]http://www.doz.com/social-media/track-social-media-kpis

[3]There is no need to consider Type I as it is very unlikely to exchange information in this case due to low physical as well as social proximity. Moreover, Type I is also a worst case scenario of Type II and Type IV

differs than Type III due to the different inherent characteristics of online social network dataset that differs from physical networks where we have limited set of information.

INDIGO addresses all cases of Physical–Social proximity table starting from Type II to Type IV. For Type II, all datasets are required to have information about the physical proximity information while for Type III case, both physical and social proximity information of people should be available. To handle social proximity between people for Type III, INDIGO considers the interests of people learned through their Smartphones or they can be created artificially using the power law distribution. For both Type II and Type III cases, INDIGO predicts the performance of data dissemination under broadcast and interest-driven data dissemination strategy respectively along with different real-world aspects. For broadcast strategy, it gives more emphasis on physical proximity information while for the interest-driven case, it models data dissemination process by contributing higher weight to interests similarity/proximity among people.

Further INDIGO also contributes in providing a solution for the Type IV (or gray area) case of Physical–Social proximity table by proposing a Machine Learning based model that predicts the performance of data dissemination on the online social network and show its effectiveness using a large-scale Twitter dataset. For Type IV, I not only provide a model to predict data dissemination but I also quantify the impact of the different set of information available on Twitter on the prediction. Finally, the Type I case modeling is not required because in this case, people are unlikely to exchange information with each other.

Therefore, the main contribution of this thesis is to provide a solution for data dissemination considering both data exchange that happens when two users physical encounters each other, up to data sharing that happens when an user disseminate a data to another user via online social link.

## 1.5 Organization of the thesis

The rest of the thesis is structured as follows:

Chapter 2 will present the State of the Art (SOA) for the efforts taken in literature for different cases of Physical–Social proximity table. I mainly present the SOA related to the heterogeneous contact patterns for both static and time-varying, multiple simultaneous contacts, different data requirements, different data dissemination strategies and multiple data sources. Further, I present the literature review not only from the data dissemination time point of view but

I also present the SOA work to find best relays in mobile networks. Finally, I also highlight the existing work to model data dissemination in online social networks.

Chapter 3 will present the INDIGO framework by providing the overview of its different modules, sub-modules, and component for all types of Physical–Social proximity table. This chapter discusses the usefulness of the modules required to predict heterogeneous mobility and the prediction of data dissemination for both broadcast and interest-driven strategy. This Chapter also presents the datasets considered in the thesis (both the standard set of datasets as well as my collected datasets) and the performance metrics predicted by INDIGO i.e. upper bound of data dissemination time and best relays in the network. Finally, this Chapter will also briefly present the part of the INDIGO framework that tackles data dissemination for online social networks.

In Chapter 4, I will focus on the Contact Probability Prediction Module of INDIGO that captures the contact and mobility patterns of people for Type II and III cases. The contact probabilities among people significantly impact the data dissemination process, therefore, the prediction of contact probabilities are important for INDIGO framework to enable the realistic prediction of the upper bound of data dissemination time. In this Chapter, I present the prediction methodologies for both static and time-varying pair-wise contact patterns. For the static contact patterns prediction I will present the Maximum Likelihood Estimation method while for time-varying contact probability prediction, I will describe the Machine Learning based increment learning that uses the Gradient Boosting Machine (GBM).

In Chapter 5, I will focus on the upper bound of data dissemination time for contact traces under broadcast data dissemination strategy for both static and time-varying pair-wise heterogeneous contact probabilities for Type II case. In this Chapter, I will present a Markov chain based model called `DDT-Markov` of INDIGO framework that can realistically predict the upper bound data dissemination time of multi-contact and multi-source data dissemination process. This Chapter also shows how `DDT-Markov` achieves much tighter and realistic upper bound of data dissemination time by utilizing the exponential cut-off property of inter-contact time distribution.

Further, in Chapter 6, I will present a more realistic aspect of data dissemination process i.e. interest-driven data dissemination strategy where people collect and share information that is interesting to them as opposed to the broadcast approach that enforces people to receive all information under Type III case. This Chapter will also present my approach to extract the interests and pair-wise interest similarities among people by utilizing the information retrieval techniques. I

also present how can we still generate artificial interests of people when interests are not available in the dataset.

Chapter 7 will focus on another dimension of data dissemination process i.e. finding the best relays in the network to speed up the information diffusion in the network by introducing the BROP model for both Type II and III cases. I proposed the methodology to find the Best Relays for broadcast and interest-driven data dissemination strategy uses the K-Shell decomposition algorithm that considers both degree centrality and links weights of each node in the network.

For all Chapters starting from 4 to 7, I validated the results obtained from each Chapter with all data sets.

In Chapter 8, I will focus on understanding and predict data dissemination using INDIGO in online social networks to handle the gray area of the Physical-Social table (Type IV) where only social proximity information is available and plays a fundamental role. In this Chapter, I will present a novel approach to predict the data dissemination based on a Gradient Boosting Machine method. I will also provide the deeper understanding of the diffusion process and quantifies the impact of the rich set of information available on online social networks by introducing different feature planes based on their complexity to acquire and privacy intrusiveness. I validated the proposed model and INDIGO framework on a large-scale Twitter dataset.

Finally, Chapter 9 summarizes the work described in this thesis with an overview of the main findings and results. The Chapter will also provide the future directions derived from this work. I have already published some of the results reported here (9 publications) and have submitted two more papers on the remaining results.

# Chapter 2

# State of the Art

## 2.1 Introduction

Mobile phones are ubiquitous devices and millions of them are used around the world. These devices are incorporated with a rich set of sensors. Due to the sensing and computational capabilities of mobile devices, they are often called smartphones. The increasing penetration of smart devices and their immense capabilities have enabled the rise and sharing of localized dissemination based services, offers, advertising where people can get them from their vicinity and further disseminate it to other people Ott et al. [2011] Guardian [2015] Google [2015]. According to these reports, people are not only interested in getting localized services but the publishers (or local businesses) of such information also need to take the right decision at the right time before pushing their offers to the interested people and maximize the reach of their information. To help publishers, there is a need to provide a unified solution to them that model and predicts the performance of data dissemination under realistic scenarios and different constraints.

In this Chapter, I will first present a set of people-centric applications in Section 2.2. Afterward, I present the literature review of data dissemination under physical proximity (in Section 2.3) where I will explain data dissemination process and also present its different real-world aspects that need to be considered for realistic performance prediction of data dissemination process. I will present the related effort done in literature for each aspect of data dissemination process and for both performance metrics i.e. data dissemination time and finding best relays. Finally, I will also present the efforts taken in literature to model and predict the performance of data dissemination in online social networks in Section 2.4.

## 2.2 People-centric localized applications

In any people-centric application, humans become the focal point of gathering and sharing any information for the benefit of each other. It empowers people to collect and share data using mobile devices across a set of applications. The concept of localized and people-centric application was initiated from the concept of Delay Tolerant or Opportunistic Networks where people rely on the device to device communication in the absence of Internet Conti et al. [2010] Pelusi et al. [2006] Jones et al. [2007]. Some of the key projects based on the concept of opportunistic networking are Haggle Scott et al. [2006b], ZebraNet Juang et al. [2002] and Diverse Outdoor Mobile Environment (DOME) Soroush et al. [2012]. In the ZebraNet project, a mixed team of biologists and computer scientists attached sensor-equipped collars to zebras in Central Kenya, with the purpose of monitoring the mobility and the social behavior of the animals. The data was to be collected and stored in the zebra collars, and finally transmitted to the researchers whenever they approached even a small subset of the animals. The system, including specialized hardware, a lightweight operating system, and a communication protocol, was designed almost from scratch, based on real-world constraints specified by the biologists. The DOME is a hybrid testbed for mobile systems, consisting of two components: DieselNet, a sparse vehicular network of public buses, and a mesh network of WiFi access points, installed on rooftops and light poles. Since 2011, DOME is publicly available for experiments, providing virtual machines on each bus, an experimenter portal for uploading experiments, as well as various control services (logging, resource allocation etc). DOME was most notably used to gain valuable insights into mobility and contact patterns in public transportation. The Haggle project defines an architecture for Opportunistic Networks formed of smart-phones, allowing both infrastructures as well as direct peer-to-peer communication. Since smartphones are carried by people, human mobility, and contact patterns, as well as social relationships are central to the Haggle architecture. Haggle experiments produced a solid understanding of human contact patterns and social relationships reflected in these contacts.

Recently, due to the proliferation of mobile devices and their integration with several sensors has further opened the door for many other advanced large-scale people-centric sensing applications Abdelzaher et al. [2007] Campbell et al. [2008]. Some of the examples of such applications are CenceMe Miluzzo and Lane [2007] which is a personal sensing system that enables members of social networks to share their sensing presence with their buddies in a secure manner. Sensing presence captures a user's status in terms of his activity (e.g., sitting, walking, meeting friends), mood (e.g., happy, sad), habits (e.g., at the gym, cof-

fee shop) and surroundings (e.g., noisy, hot). Once the application collects the sensing presence of a person, it injects this information into the popular social networking applications such as Facebook, MySpace, and IM (Skype, Pidgin) allowing for new levels of connection and implicit communication between friends in social networks. The another application is CycleSense Mun et al. [2009] which is a web-based application designed to give bike commuters feedback on the quality and safety of their preferred routes and to suggest quality-of-ride in a particular route. Bike commuters use their mobile devices to gather information of their biking routes using accelerometer and GPS sensors. Further, they also share their personal feedback to the particular route in terms of ease of riding, meeting with other bikers etc. Finally, they upload this information to the central sever to serve as the recommendation for other interested bikers. Similar to this application BikeNet is another application that maps the cyclist experience Eisenman et al. [2009].

The application called Twimight is developed to be used in the disaster situations when infrastructure is hardly available or completely wiped off Hossmann et al. [2011a]. Twimight behaves like a normal Twitter client but when it is set to "disaster mode", it enables opportunistic communication where tweets spread epidemically among the people to share the recent information. FireChat is another example of a people-centric message application that share and disseminate localized information to the different set of people Gardern [2015]. This application is a proprietary mobile application developed by OpenGarden and has been successfully deployed during natural disasters including floods in Kashmir (April 2015) and Chennai (October 2015), the eruption of volcano Cotopaxi in Ecuador (August 2015), and hurricane Patricia in Mexico (October 2015) and massive events like pro-democracy protests in Taiwan (April 2014), Hong Kong (September 2014), and the visit of the Pope in the Philippines (January 2015) and large festivals in India, Canada and the US. With the help of this application, people do share information about their local environment. The sense of localized information is also adopted in the Walt Disney World Resort in Florida to study the mobility of people and to enable the distribution of park information (waiting times at different attractions, schedules of street parades and other performances), mobile advertising, participatory sensing, polling/surveying, and multimedia sharing Vukadinovic and Mangold [2011]. The project called PROMO (PROximity Marketing sOlution) is one of the recent and relevant applications that aim to exploit local knowledge Papandrea et al. [2010] Vanini et al. [2012]. PROMO enables users to receive mobile advertisements depending on their current location and specific interests. A user connects his/her smartphone to nearby Wi-Fi access point and receives offers (and advertisements) from

shops that are within a certain range. These advertisements are stored in a central server and transferred to the user through different Wi-Fi access points. Recently Qualcomm is also introducing the concept of innovative LTE direct device-to-device technology that enables mobile devices and apps to passively discover and interact with the world around them in a privacy sensitive and battery efficient manner. Implementation of the LTE Direct ecosystem is underway and it is going to give a rise to new proximity service opportunities for the entire mobile industry in social networking, venue services, loyalty services, local advertising, and much more. Qualcomm [2014].

Finally, apart from physical networks, the localized information like local offers, advertisements, events, new products are also shared on the online social networks like Twitter, Facebook, Instagram etc. These social networks also serve as people-centric applications under the avialability of social proximity where people do gather and share relevant information with each other online. These sites allow advertisers to identify new topics that are gaining interest or "trending" rapidly across the platform Du and Kamakura [2012]. Further, these online social networking sites also allow advertisers to identify users who propagate these newly trending topics, and to target advertising specifically to them Vaynerchuk [2013]. The marketers often try to seed information about their product or service with such users, hoping they will engage with it and spread it virally to their peers.

## 2.3 Data dissemination under physical proximity

### 2.3.1 What is data dissemination process?

Usually, in data dissemination process, people exchange data when they are in physical proximity (in contact) Boldrini and Passarella [2013]. These people store and carry data through their mobility and eventually forward it to others, thus achieving multi-hop communication despite the lack of end-to-end paths. The easiest and straightforward process to disseminate data is known as *Epidemic Spreading*. It operates as follows: given a piece of information/message $m$, every node carrying a copy of $m$ must further replicate the message to every node it encounters (provided the encountered node does not already have $m$). Thus, information will spread almost like an epidemy through the network, with every node eventually receiving a copy of $m$. Epidemic spreading is mainly studied in disease spreading Yoneki [2011] and Delay Tolerant Networks Zhang [2006]. It is important to note that this basic data dissemination process does not take into

account data freshness and interests of people during data exchange.

### 2.3.2 Real-world aspects of data dissemination process

Figure 2.1 presents key aspects of data dissemination process that need to be considered while predicting the data dissemination performance in a mobile scenario for Type II and III cases of the INDIGO data dissemination framework. The key aspects (shown in rectangular boxes) are contact patterns among people, communication among people, different data sources, different data dissemination strategies, and data requirement of people. Further, I also present specific topics of these key aspects and represent them through oval boxes. The topics that have been extensively explored in literature are marked with green color while less-studied topics that need to be considered to mimic real-world aspects of data dissemination process are marked with red color. I will now present the work done in literature for each real-world aspect of data dissemination process.

#### Contact/Mobility patterns of people

The contact patterns of people significantly impact the rate of data dissemination. If people meet each other frequently then the data dissemination will be faster and vice versa. Existing works model the spread of information either under different mobility models like Random Mobility, Random Waypoint etc. Clementi et al. [2012] or homogeneous pair-wise contact rate among people Groenevelt et al. [2005]. However, these random movement based mobility models or homogeneous contact rates among people do not represent the real mobility and contact patterns of people. Therefore, these models fail to provide a realistic evaluation of data dissemination process.

The work done in Conan et al. [2007] Lee et al. [2009] Passarella and Conti [2013] have emphasized to utilize an individual level (pair-wise) distributions of inter-contact time among people rather than assuming aggregated distribution of inter-contact time. Further, recent experimental studies Eagle et al. [2009] Foerster et al. [2012] also show the existence of considerable heterogeneity in node mobility, thus questions the predictions of existing models. The work done by Picu et al. [2012] Boldrini et al. [2014] have considered fixed pair-wise heterogeneous contact rates for the entire duration of time among people and provides an estimation for the bounds of data dissemination time in opportunistic networks. However, in a real-world scenario contact patterns of people vary with time and context, therefore, we cannot assume fixed pair-wise heterogeneous contact rates for the entire duration. I also presented the presence and signifi-

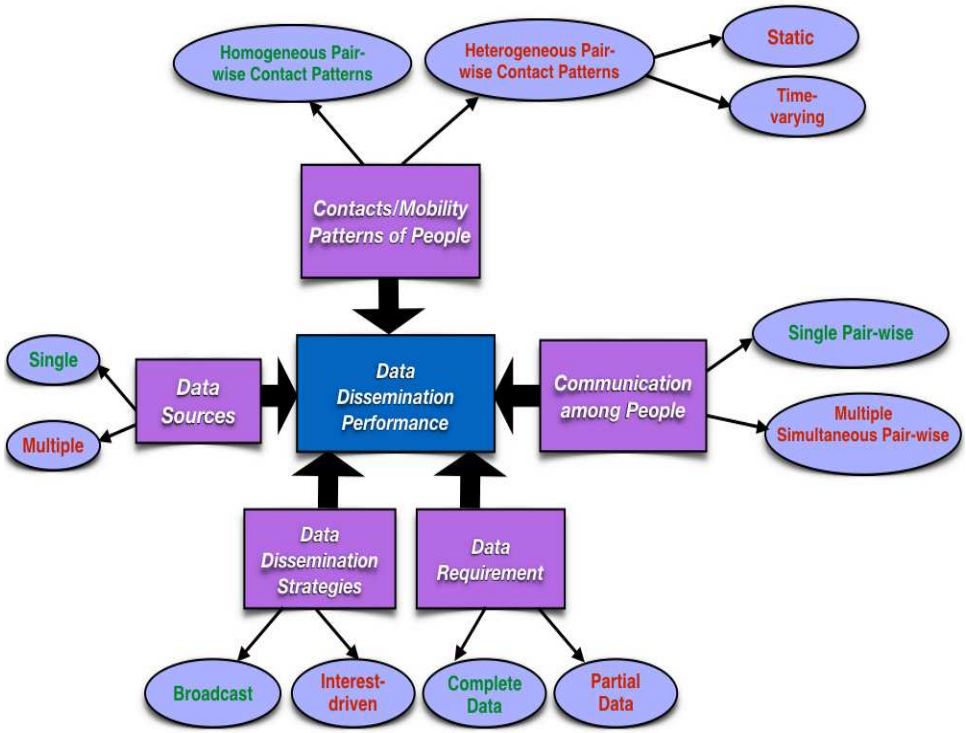



Figure 2.1. Different real-world aspects required for the performance prediction of data dissemination process in physical networks. A square box represents a broad aspect and an oval box represents a specific topic. The green color represents significantly studied topic and red represents a topic that aligned with real-world scenario.

cance of heterogeneous and time-varying contact patterns in Chapter 1. Some of the existing works also consider heterogeneity in terms of space Lee et al. [2009] and community Hui et al. [2008] Hossmann et al. [2011b]. They assume that people who are residing in the same geo or social community have same contact rates and should be able to exchange message with them. These works are the specific case of heterogeneous contact rates where some pair of people has same contact rate. However, INDIGO further dive down and consider the pair-wise contact patterns to identify individual-level contact patterns to enable much better prediction of data dissemination process.

Further, the time-varying heterogeneous contact probability prediction is still not addressed in literature even though it represents more realistic aspects of human mobility. The work done in Gao et al. [2013] has shown that people show transient contact patterns in different time slots in a day which does not necessarily be symmetric and can be different for different days. The authors in Casteigts et al. [2012] have tried to model a mobile network as the dynamic time-varying graphs with several snapshots at different time intervals. However, the paper requires complete link information about the network to create different snapshots of the network and also rely on manually finding the time slot of the time-varying graph. The another work done in this direction predicts the link quality based on the signal to noise ratio in mesh networks using the matching techniques Farkas et al. [2008]. Recent works have also shown that over a longer duration, people do exhibit regularity in their mobility patterns Hsu et al. [2009] Gonzalez et al. [2008] and also show that link prediction can be possible in context to social and location based networks utilizing Machine Learning approaches Liben-Nowell and Kleinberg [2007] Scellato et al. [2011] Hawelka et al. [2015]. The Machine Leaning models are extensively used in social networks, location–based traces however they are still not widely employed for contact traces. The work done by Jahanbakhsh et al. [2012] is the first work that has used a supervised Machine Learning approaches to find the hidden contacts among people which were not captured during the experiment. Therefore this work still does not focus on predicting the future contact patterns among people that vary over time.

### Communication among people

In a real-world scenario, at any given time multiple pairs of people can exchange information with each other and also impacts the performance of data dissemination. Therefore, while predicting data dissemination performance, we also need to model multiple simultaneous contacts among people. A majority of

works consider that at any given time, at most only one pair of people can communicate with each other thus, discount the impact of multiple simultaneous contacts on data dissemination performance Picu et al. [2012] Groenevelt et al. [2005] Clementi et al. [2012]. To the best of my knowledge, this topic is still not well studied in the literature while modeling data dissemination process in mobile networks.

### Data dissemination strategies

In real-world, people usually exchange information based on their shared interests, context and freshness (time validity of data) and do not simply flood information Boldrini and Passarella [2013]. Therefore, to predict realistic performance of data dissemination, the consideration of different data dissemination strategies is essential as it decides how data actually gets disseminated in a given scenario. Most of the existing works utilize a simple broadcast strategy while predicting the performance of data dissemination process Groenevelt et al. [2005] Mosk-Aoyama and Shah [2008] Clementi et al. [2012].

The broadcast strategy also comes under gossip and epidemic protocols. The gossip protocols are mainly used in a computer–to–computer communication protocol inspired by the form of gossip seen in social networks Jelasity et al. [2005]. Further, epidemic protocol models virus propagation to understand the spread of disease among people Yoneki [2011]. It is also used in the context of Delay Tolerant Networks to understand information propagation Zhang [2006]. Both Gossip and Epidemic protocols utilize broadcast strategy to exchange gossip or virus/information through random selection of people. The another stream of data dissemination strategies is the Peer-to-Peer (P2P) Systems that are operated in a completely decentralized fashion that means there is only peer-to-peer data exchange. In this strategy, there is no scope to get data directly from the server. Also, most of the P2P techniques for data dissemination do not consider interests of people and other real-world mobility aspects like multiple simultaneous contacts among the different pair of peers Zhou et al. [2011].

INDIGO differs from these protocols due to the consideration of multiple simultaneous contacts among people, multiple data sources, data requirements for broadcast strategy while it also considers the interests similarities among people. The Gossip, Epidemic protocols and Peer-to-Peer (P2P) Systems do not consider these real-world aspects. Further, these strategies only rely on getting data from peers. In the case of INDIGO, I consider heterogeneous communications i.e. data can either be received from the central server (directly from the publisher) or it can be further augmented with the help of peer-to-peer data dissemination pro-

cess Boldrini and Passarella [2013].

With respect to interest-driven data dissemination, the most closer work that has been studied in literature is the Publish-Subscribe schemes where some node (or people) that generate content are termed as publishers, while nodes subscribing to that content are termed as subscribers. Further, there is also a set of brokers that serve as mediators to provide a publisher's content to subscribers. Brokers have complete knowledge of all subscriptions and the content generated by all publishers Boldrini et al. [2010] Mei et al. [2011] Boldrini and Passarella [2013] Ciobanu et al. [2015]. In these methods, users explicitly show their interests in certain topics. Asking user interests are neither feasible in long term nor scalable because it limits the wide range of interests user can express and their validity over a long time. Further, these works do not enable the automatic learning of user interests. INDIGO differs from pub/sub systems because people do not subscribe their interests to publishers. The interests of people and their importance are learned automatically and updated locally and data is only exchanged if interests similarity between two people is high. Further, INDIGO also do not explicitly select brokers for data exchange however, it is able to detect a set of best people for publishers to enable faster data dissemination process using BROP Model. The authors in Zhou et al. [2013] also focus on the incentive-based data dissemination technique where people exchange information once they receive some incentive which could be quite useful. This thesis currently does not focus on such aspect of data dissemination process.

To conclude, in this thesis, I focus on two types of popular data dissemination strategies: broadcast and interest-driven for both Type II and III cases respectively. In broadcast strategy, people exchange data only based on their physical proximity however in the case of interest-driven strategy, data is also exchanged based on the interests of people where the interests of people are learned automatically. To the best of my knowledge, in literature interest-driven data dissemination along with the automatic retrieval of interests is not studied for the physical networks where contact traces are prominent. Therefore, this thesis fills this gap not only though the learning and modeling of interest-driven data dissemination strategy but it also gather the first traces that consists the contacts of people along with their interests depicted through their on-mobile web browsing history.

### Data requirement

People have different data requirements as some people require to collect only a fraction of data while others require complete data from the network. For ex-

ample, information about the parking place is enough for person A, while for person B information about nearest shopping place is also important. In literature, most of the work evaluated the performance of data dissemination for the complete data collection or the maximum amount of data that can be collected in the network Picu et al. [2012] Boldrini et al. [2014]. In the real-world scenario, we need to consider different data requirements where people are interested in the certain fraction of the information. More specifically this case arises for the interest-driven data dissemination strategy where data is disseminated solely on the interests of people thus leads to fewer data collection. Once again, INDIGO fills this gap in the literature by considering different data requirements of people while predicting the performance of data dissemination.

#### Data sources

Generally, in a real-world scenario, data can originate from multiple sources. If we consider our touristic city scenario, people can receive information from multiple data sources like parking information, event information, advertisements etc. These data sources send information at different time intervals. Existing works only consider single data source while evaluating the performance of data dissemination process Groenevelt et al. [2005] Picu et al. [2012] Boldrini et al. [2014]. In this thesis, I take into account the impact of multiple data sources while predicting the performance of data dissemination process.

From the above literature review, I find that most of the existing work focus on single-$*$ scenarios, however, in order to realistically predict data dissemination performance, we need to *collectively* consider these real-world aspects. INDIGO fills this gap by providing a generalized data dissemination framework that enables realistic performance of data dissemination under real-world mobility, communication and strategy aspects.

### 2.3.3 Modeling of data dissemination process

The modeling of data dissemination has been first studied theoretically in literature. The authors of Pettarin et al. [2011] and Peres et al. [2011] present data dissemination bounds using percolation theory by considering the network as a dynamic graph. The theoretical results for the upper bound of broadcast time in context to an information spreading algorithm is presented in Mosk-Aoyama and Shah [2008] using separable functions by considering both synchronous and asynchronous model and multiple data sources. The work done in Picu and Spyropoulos [2010] analytically studies the broadcast time for data dissemination

under multiple data sources using coupon distribution problem Feller [2008]. Some works have modeled the dissemination process as fluid flows using ordinary differential equations (ODEs) Zhang et al. [2007], where the number of copies in the network is approximated by a continuous-valued function of time and node meeting rate. These theoretical results provide an extremely high overestimation of broadcast time thus, questions their applicability in the real world environment.

The existing works also model data dissemination process using Markov chains under random mobility models like Random Mobility, Random Waypoint etc. Groenevelt et al. [2005] Clementi et al. [2012] where each Markov state also represents a number of copies in the network. The above studies consider node mobility as stochastic and independent identically distributed (IID) process with homogeneous node meeting rate $\lambda$. However, recent studies Eagle et al. [2009] Foerster et al. [2012] show that there exists considerable heterogeneity in node mobility, thus questioning the predictions of existing models. An excellent piece of work done by authors in Picu et al. [2012] consider heterogeneous node contact rates and model single source data dissemination process using Markov chains. The similar work is also done in Boldrini et al. [2014] where the data dissemination is once again modeled using Markov chain and focusing on the importance of pairwise heterogeneous contact patterns of people. The above works neither model time-varying contact patterns of people nor do they consider other real world mobility aspects such as multiple simultaneous contacts among people and modeling of their interests. Further, all of the above these models only predicts the performance of data dissemination from the perspective of data dissemination time. They also fail to provide tighter upper bound of data dissemination time. INDIGO is able to model different aspects of mobility, communication and data dissemination and provide a unified solution by combining different methods along with Markov chain models.

The other performance prediction dimension that I am considering in this thesis is finding the best relays in the mobile network to accelerate the data dissemination process. Searching for best spreaders in complex networks is studies across various domains, ranging from the epidemic control Anderson et al. [1992] Heesterbeek [2000] Pastor-Satorras and Vespignani [2001], viral marketing Watts et al. [2007] Leskovec et al. [2007] and social movement to idea propagation Diani and McAdam [2003] Lü et al. [2011] Myers et al. [2012b] Zhang et al. [2016]. To find the super spreaders in such complex networks, these works focuses on the network properties using centrality measures like degree centrality (or just the degree of a node, i.e. the number of its links), the eigenvector centrality Bonacich [1987], the betweenness centrality Freeman [1977] etc. Re-

cently, the another centrality measure based on the notion of K-cores is applied in many real networks Dorogovtsev et al. [2006] Carmi et al. [2007] Garas et al. [2010] Basaras et al. [2013] Pei et al. [2014] and shown to be effective in understanding network structure and finding influential nodes in the network Batagelj and Zaveršnik [2011].

One of the major limitation of the above-described centrality measures, including the K-core decomposition method, is their design to work on unweighted graphs. However, in practice, real networks are weighted that describe important and well-defined properties of the graph nodes. To handle such complex networks, the authors in Garas et al. [2012] proposed a weighted K-Shell decomposition algorithm that takes into account both the degree centrality and weight measures to find best relays in the network. In the case of INDIGO, I also utilize the weighted k-Shell decomposition algorithm and propose a methodology to model contact strength (physical proximity) and interest similarity (social proximity) among people as a weighted graph and address the problem of finding best relays in the network that minimizes the data dissemination time.

I summarize different work done in the direction of data dissemination performance prediction and evaluation metrics considered in Tables 2.1 and 2.2 respectively.

From Tables 2.1 and 2.2, I conclude that existing works do not collectively consider multiple real-world aspects of data dissemination process and only predicts the bound of data dissemination time. INDIGO provides a unified solution and fills the gap of literature by collectively considering real-world aspects of data dissemination and predicts its performance in multiple dimensions.

## 2.4 Data dissemination under online social proximity

The data dissemination can also occur in the case of the availability of social proximity information in online social networks where people acquire information and influence each other based on their friendship and interests Chen et al. [2013]. In this section, I will present the research efforts taken to disseminate information in online social networks (Type IV) of the INDIGO framework.

Social networks have been studied extensively by social scientists and were confined to small datasets. Enabled by the Internet and sparked by the recent advancement of online social networking sites such as Facebook, Twitter and LinkdIn, research on social networks is witnessing an unprecedented growth due to the ready availability large-scale social network data. This has also led to the development of several applications and opened the door for different research

Table 2.1. Comparison of state of the art with respect to different data dissemination aspects.

| | Heterogeneous Contact Patterns | | Communication Among People | | Data Diss. Strategies | | Data Sources | | Data Requirement | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Static | Time-Varying | Single Pair-wise | Multiple Pair-wise | Broad-cast | Interest-driven | Single | Multi | Full | Partial |
| Picu et al. [2010] | ✓ | × | ✓ | × | ✓ | × | ✓ | ✓ | ✓ | × |
| Pettarin et al. [2011] | ✓ | × | ✓ | × | ✓ | × | ✓ | × | ✓ | × |
| Clementi et al. [2012] | ✓ | × | ✓ | × | ✓ | × | ✓ | × | ✓ | × |
| Picu et al. [2012] | ✓ | × | ✓ | × | ✓ | × | ✓ | × | ✓ | × |
| Boldrini et al. [2014] | ✓ | × | ✓ | × | ✓ | × | ✓ | × | ✓ | × |
| INDIGO | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

directions for both businesses and researchers. A rich body of such research is classified as the analysis of information propagation (or dissemination) in online social networks. Typically this phenomenal is realized by the means of like, post/message sharing and retweet on social networks like Facebook and Twitter. A sequence of posts/retweets along the network is called information cascade. Since in this thesis, I am focusing on the Twitter dataset, therefore, I will present the literature review related to data dissemination in Twitter.

Although some initial work has been done to model complete diffusion cascades in social media Galuba et al. [2010]; Petrovic et al. [2011], researchers have recently argued that cascades might be inherently unpredictable, due to the high number of factors, either internal or external to the network Myers et al. [2012a], that affect the outcome of diffusion Salganik et al. [2006]; Martin et al. [2016]. For this reason, predicting the exact pattern of diffusion of a piece of information starting from a given node in the network remains challenging.

While modeling data dissemination in online social networks, the real world dissemination aspects are different as compared to the physical networks as the

Table 2.2. Comparison of state of the art with respect to different data dissemination evaluation metrics.

| | Performance Metrics | | Validation |
|---|---|---|---|
| | Data Dissemination Time | Best Relays | With Real Traces |
| Picu et al.[2010] | ✓ | × | ✓ |
| Pettarin et al.[2011] | ✓ | × | × |
| Clementi et al.[2012] | ✓ | × | × |
| Picu et al.[2012] | ✓ | × | ✓ |
| Boldrini et al.[2014] | ✓ | × | ✓ |
| INDIGO | ✓ | ✓ | ✓ |

datasets associated to online social networks differ in characteristics in terms of different type of rich information. In case of Type IV case, the real-world data dissemination aspects I considered are: the heterogeneous activity of different users on online social networks, type of information that needs to be disseminated, friendship ties, associated groups/communities and the content of the published online activities.

Most of the works in literature are mainly focusing on the analysis of specific aspects of information diffusion in social networks, such as whether diffusion will grow in future or not Cheng et al. [2014], the impact of content sentiment on diffusion Ferrara and Yang [2015], and the effect of features related to items or users on content popularity Hoang and Lim [2012]; Yang and Counts [2010]. The work by Yuan and colleagues Yuan et al. [2016] is focused on the impact of social relationships and tie strength on the probability of diffusion. The work aims at sorting the friends of a user by their likelihood to retweet or reply its tweets and, does not specifically address information diffusion. The authors in Pezzoni et al. [2013] analyzed the impact of temporal features and popularity indicators on the diffusion. The results indicate that content age and its visibility in the homepage of the user strongly influence the probability of resharing.

Another research area related to the analysis of single step diffusion is from the perspective of personalized tweet recommendation. This approach aims to recommend tweets that could be interesting to the users instead of predicting whether users will reshare them in the future. On this line of research, Chen et al. use several features related to users profiles and their similarity, the con-

tent of tweets, and the social relationships between users to recommend existing tweets to users Chen et al. [2012]. A similar solution is also proposed by Hong et al. Hong et al. [2012].

The above works model information diffusion by only predicting the retweet probability while utilizing some of the features available in the online social network. As opposed to other works, INDIGO framework also take into account reply for information propagation by predicting the likelihood to reply to a particular tweet. Since a rich set of information is available on Twitter starting from users profile to content analysis, therefore, in this thesis, I not only provide a way to predict data dissemination by predicting the likelihood of retweet and reply but also quantify the importance of different information by introducing the concept of feature planes. Existing works in this area mainly tried to predict information spread by utilizing specific aspects of information like social network structure, temporal properties, profile features and topical features Galuba et al. [2010]; Petrovic et al. [2011]; Yang and Counts [2010]; Pezzoni et al. [2013] but none of them successfully combined all these features together and, more importantly, they do not quantify the importance of different features for retweet prediction. I argue that a fundamental knowledge of different feature planes (defined as a group of features with similar cost in terms of privacy and complexity to acquire), their individual and combined contribution in retweet prediction has to be analyzed for better prediction of information diffusion. In this way, I not only address the problem of data dissemination for online social networks but also enable to reduce the complexity of the model by providing the trade-off between high prediction accuracy and privacy. Also, the proposed model of INDIGO framework does not limit the prediction of retweet and reply to tweets that are generated by the friends of the user rather predicts it for any generic tweet.

## 2.5 Summary

After describing existing works in the different aspects and types of data dissemination process, it is evident the need for a framework that collectively considers multiple real-world aspects such as heterogeneous and time-varying contact patterns, interests of people, different data dissemination strategies, multiple-simultaneous contacts among people, data originating from multiple sources etc. for different cases of physical and social proximity. In the next chapters, I will build on such works to propose my novel data dissemination framework INDIGO that aims to overcome the limitations highlighted in this Chapter.

# Chapter 3

# Overview of INDIGO: A Generalized Data Dissemination Framework

## 3.1 Introduction

In this Chapter, I will present different parts and working of the INDIGO framework to predict the performance of data dissemination process for Type II, Type III and Type IV cases of the Physical–Social proximity table. As presented in Figure 3.1, INDIGO addresses the availability of both physical and social proximity information among people for Type II, Type III using the *INDIGO–Physical* part for the diverse environments by collectively considering the realistic mobility, communication and social aspects of people. In both cases, all datasets are required to have information about the physical proximity of people[1]. Since the social proximity information is not available for For Type II case, therefore, INDIGO models this case under broadcast data dissemination strategy. Further for Type III case, INDIGO models the social proximity between people through the interests of people learned through their Smartphones and predicts the performance of data dissemination under interest-driven data dissemination strategy. For broadcast strategy, it gives emphasis on physical proximity while for the interest-driven case, it models data dissemination process by contributing higher weight to interests similarity/proximity among people. Finally, the *INDIGO–OSN* part of the INDIGO framework presents the efforts taken in this thesis for the gray area i.e. Type IV where social proximity among people plays the fundamental role.

The chapter is structured as follows. In Section 3.2, I will present the brief overview and working of different modules, sub-modules, and components of IN-

[1]Type I modeling is not required because in this case, people are unlikely to exchange information.

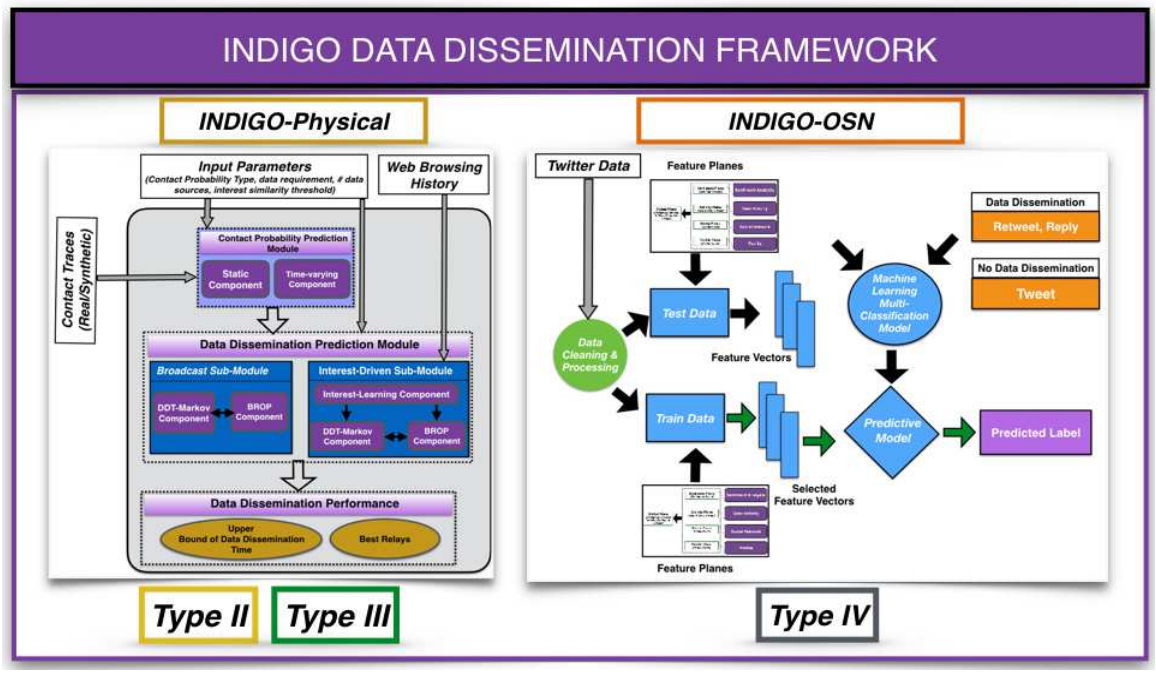



Figure 3.1. INDIGO data dissemination framework that predicts the performance of data dissemination for both physical and online social networks. The first part (physical networks) of the INDIGO handles Type II and III case of the Physical-Social proximity table while the second part (online social networks) handles the gray area i.e. Type IV case of the Physical-Social proximity table.

DIGO for the Type II and III cases. The section mainly cover information about the dataset considered in this thesis (Section 3.2.1), modules required to predict heterogeneous mobility (Section 3.2.2), prediction of data dissemination (Section 3.2.3) for both broadcast and interest-driven strategy along with the input parameters given to INDIGO framework. This Section also presents the two performance metrics predicted by INDIGO i.e. upper bound of data dissemination time and best relays in the network (Section 3.2.6). Section 3.3 will present the different parts of the INDIGO framework required to predict the performance of data dissemination for Type IV case. Finally, Section 3.4 concludes the Chapter.

## 3.2 Overview of INDIGO for Type II and Type III cases

Figure 3.2 outlines different modules, sub-modules, components and sub-components of the *INDIGO–Physical* part of INDIGO framework required to model Type II

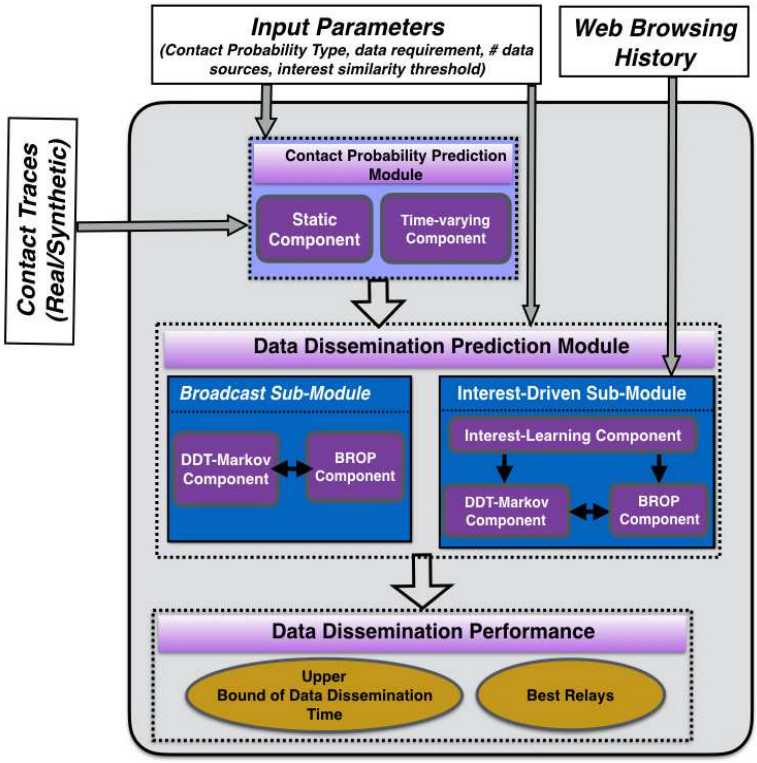



Figure 3.2. Overview of the different components and their working of INDIGO framework for Type II and III cases where physical proximity plays the important role.

and III cases where physical proximity plays the important role. I will provide the brief overview and working of the each part along with the description of dataset and inputs required to model data dissemination process.

### 3.2.1 Contact traces

To model and predict the performance of data dissemination, INDIGO takes either the synthetic or real contact traces. From these traces, it predicts the contact patterns among people and utilizes it for both Type II and Type III cases to model real-world mobility aspects of people. In this thesis, I consider some standard and most widely set of contact traces from diverse environment used in literature i.e. INFOCOM, MIT, and ROLLERNET. Further, during my Ph.D., I also collected two set of traces PERCOM and MACACO from conference and university environment by conducting my own experiments. I will now briefly describe each contact traces used in this thesis.

### INFOCOM

This contact trace was collected during a conference as a part of the Haggle project, one of the major projects in the field of opportunistic networking Haggle [2006]. During Haggle experiments, people were asked to carry an experimental device called Intel iMotes[2] with them at all times Chaintreau et al. [2007]. These devices log all contacts between experimental devices using a periodic scanning every 120 seconds via Bluetooth. Each contact is represented by a tuple (MAC address, start time, and end time). The experiment was conducted during the IEEE INFOCOM 2005 conference in Miami, where iMotes were carried by 41 attendees for 4 days. I downloaded the contact data of INFOCOM from CRAWDAD archive where the anonymized version of the data was available Scott et al. [2009].

### PERCOM

The PERCOM was traces was collected by me during the PerCom 2012 in Lugano, Switzerland SCAMPI [2012] via Bluetooth discovery. To collect these traces, I co-developed an Android mobile application called bCards that allows exchanging digital business cards among people [3]. The digital business cards consist participant's professional details like name, affiliation, designation etc. Figure 3.3 presents screenshots of the bCards mobile application deployed to collect PERCOM traces. Over 55 people participated in data collection by downloading the application from Google Play Store from different parts of the world. All the participants collected Bluetooth contacts with their peers for the duration of 5 days. The data was stored on the secure server of SUPSI. The participants volunteered to become part of the data collection in exchange for a smartphone prize at the end of the conference.

### ROLLERNET

The RollerNet experiment conducted to collect contact traces for urban environment in Paris, France during rollerblading on August 20, 2006. It adopts the classic approach of logging contacts between participants of the roller tour. The total duration of the tour was about three hours, composed of two sessions of 80 minutes, interspersed with a break of 20 minutes. The contacts were logged on 62 volunteers using iMotes and cell phones. Participants with cell phones

[2]http://wsn.cse.wustl.edu/images/e/e3/Imote2_Datasheet.pdf

[3]The application is available at Google Play Store
https://play.google.com/store/apps/details?id=supsi.dti.percom&hl=en

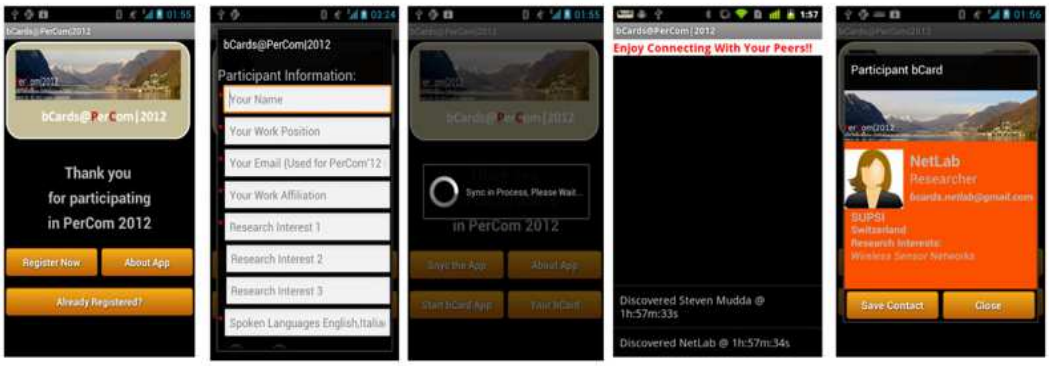



Figure 3.3. Screenshots of bCards Application deployed during PerCom 2012 to collect contact data.

were asked to activate Bluetooth on their cell phones. The iMotes also use Bluetooth technology and log periodically (every 15 seconds) the encounters they have with other devices (iMotes or cell phones) RollerNet [2006] Tournoux et al. [2009]. I downloaded the ROLLERNET data from CRAWDAD Benbadis and Leguay [2009].

### MIT

The MIT data was collected from the University environment as the part of the Reality Mining project conducted from 2004–2005 at the MIT Media Laboratory Eagle and Pentland [2006] Eagle et al. [2009]. The Reality Mining study followed 94 subjects using mobile phones pre-installed with several pieces of software that recorded and sent the researcher data about call logs, Bluetooth devices in proximity of approximately five meters, cell tower IDs, application usage, and phone status. Subjects were observed using these measurements over the course of nine months between September 2004 and June 2005. The 94 subjects of trace included students and faculty from two programs within a major research institution. Out of these 94 subjects, 68 were colleagues working in the same building on campus (90% graduate students, 10% staff) while the remaining 26 subjects were incoming students at the university's business school. The subjects volunteered to become part of the experiment in exchange for the use of a high-end smartphone for the duration of the study. The Bluetooth contacts among subject were logged with the scanning period of 300 secs. I downloaded this MIT contact data from Reality Mining project Nathan Eagle [2006].

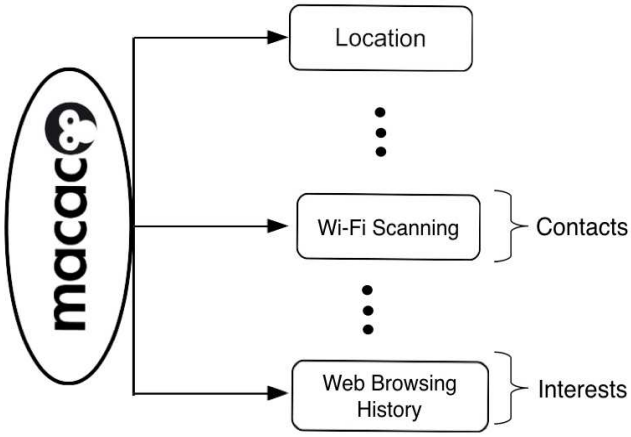



Figure 3.4. Different sensor data collected using MACACO mobile application.

MACACO

The contact traces described above do not contain information about the social proximity of people. However, to investigate the accuracy of INDIGO for Type III case, I require real-world traces that consists both contact patterns and interests of people. To the best of our knowledge, there are no real-world traces that contain this type of data, therefore, to fill this gap, I collected our own data by utilizing our own dedicated mobile application developed as a part of CHIST-ERA MACACO project in collaboration with project partners in France and Brazil MACACO [2012]. The application is also available at Google Play Store [4]. The application collect information from different sensors like GPS, accelerometer. Wi-Fi scanning etc. In addition to this, the MACACO application also logs the web browsing history on user's smartphone. From this data, I use Wi-Fi scanning data to create contacts among people (physical proximity) and browsing history of people to reflect their interests (social proximity) as shown in Figure 3.4.

Our mobile application periodically scans Wi-Fi access points connected to smartphones of users with a sampling frequency of 5 minutes. To ensure the privacy of users, our app anonymizes the identity of users and sends collected data to a secure central server. We deployed our mobile application on 27 volunteers residing in two different countries France (from April 2015 to May 2015) and Brazil (from September 2015 to October 2015). Most of the volunteers were students and staff members of universities whose mobility and interests are captured through our app that runs in the background of their smartphones. From

[4]https://play.google.com/store/apps/details?id=fr.inria.macaco&hl=en

the collected data, I create contact traces by using Wi-Fi scan information in a manner similar to Conan et al. [2006] Hsu et al. [2007]. I utilize users sessions in wireless networks by analyzing the time at which a user associates or dissociates from an access point. Finally, I create a contact between any two volunteers if they are associated with the same access point within a duration of 30 minutes. Finally, Table 3.1 summarizes the characteristics of all dataset used in my thesis to evaluate the applicability and accuracy of INDIGO for Type II and III cases of the Physical–Social proximity table using *INDIGO–Physical*.

Table 3.1. Datasets characteristics.

| | **INFOCOM** | **PERCOM** | **ROLLERNET** | **MIT** | **MACACO-France** | **MACACO-Brazil** |
|---|---|---|---|---|---|---|
| Context & Environment | Conference attendees | Conference attendees | Tourists and volunteers | Campus students and staff | Students and staff of different departments | Staff members of same department |
| Participants | 41 | 37 | 62 | 92 | 19 | 8 |
| Time Span | 3 days | 3 days | 3 hours | 9 months | 4 Weeks | 3 Weeks |
| Scanning Interval | 120s (Bluetooth) | 60s (Bluetooth) | 15s (Bluetooth) | 300s (Bluetooth) | 300s (Bluetooth) | 300s (Bluetooth) |
| Number of Contacts | 22459 | 66244 | 89498 | 81961 | 39786 | 26392 |

### 3.2.2 Contact Probability Prediction Module

The *Contact Probability Prediction Module* is responsible for predicting the heterogeneous pair-wise contact probabilities among people utilizing the real/synthetic contact traces under both static and time-varying contact patterns. Based on the *Contact Probability Type* input parameter (0 for static contact probabilities and 1 for time-varying contact probabilities), it decides to predict heterogeneous pair-wise contact probabilities for static or time-varying contact patterns using it's `Static` or `Time-varying` Components respectively. This module drives the physical proximity among people for both Type-II and III cases.

#### Static Component

The `Static Component` enables the prediction of contact probabilities among people while considering static pair-wise contact patterns. In this case, it assumes that each pair of people will exhibit same contact patterns during different

time periods. To predict the static pair-wise heterogeneous contact probabilities among people, the `Static Component` first finds the distribution of inter-contact time for different contact traces followed by employing the Maximum Likelihood Estimation (MLE) method Scholz [1985]. Using MLE method, it constructs a contact probability matrix between each pair of node and further provides this input to the *Data Dissemination Prediction Module* of INDIGO.

#### Time-varying Component

The `Time-varying Component` empowers INDIGO to model time-varying contact patterns of people by predicting the pair-wise heterogeneous time-varying contact probabilities among people. Since the assumption of static contact probabilities do not hold in reality as people do not meet with each other with same likelihood thus the `Static Component` cannot capture varying mobility patterns of people over time. As a result, it does not provide the realistic prediction of data dissemination process. To predict time-varying contact probabilities, the `Time-varying Component` employs a Machine Learning approach that learns the contact pattern of people over time and generates a contact probabilities matrix set for different time slots and input it to the *Data Dissemination Prediction Module* of INDIGO framework. However, there are situations where time-varying cannot be used due to its requirement of the longer dataset. In those cases, INDIGO utilizes the `Static` component. More details about the *Contact Probability Prediction Module* and its components will be presented in Chapter 4.

### 3.2.3 Data Dissemination Prediction Module

It is the core module of the INDIGO framework to predict the performance of data dissemination time in two dimensions: tighter upper bound of data dissemination time and best relays in the network that allows disseminating information quickly. It receives input from *Contact Probability Prediction Module* along with other Input Parameters. The *Data Dissemination Prediction Module* consists of two sub-modules `Broadcast Sub-Module` and `Interest-Driven Sub-Module`. The `Broadcast Sub-Module` is responsible for predicting data dissemination performance under broadcast strategy (Type II) while the `Interest-Driven Sub-Module` addresses Type III case by considering the interest-driven data dissemination technique. Both sub-modules predict data dissemination performance for either static or time-varying contact patterns.

#### Broadcast Sub-Module

This sub-module is responsible for predicting the tighter upper bound of data dissemination time and best relays in the network for broadcast dissemination strategy using *DDT-Markov* and *BROP* components respectively. The *DDT-Markov Component* further consists of Data Processor, Markov Model, and Cut-off Estimator sub-components to predict upper bound of data dissemination time. The Data Processor pre-processes different inputs required by the Markov Model that employs a Markov-chain based model for both static and time-varying contact patterns. The Cut-off Estimator is responsible for providing the tighter bound of dissemination time by estimating a Cut-off point $\alpha$ (details of $\alpha$ is provided in next Chapters). Finally, the *BROP* component is responsible for finding the best relays in the network under different broadcast dissemination strategies. To do this, it utilizes a weighted k-shell decomposition algorithm Garas et al. [2012].

#### Interest-Driven Sub-Module

The `Interest-Driven Sub-Module` predicts the tighter upper bound of data dissemination time and best relays in the network under interest-driven dissemination strategy and both contact patterns. Similar to `Broadcast Sub-Module`, it also consists *DDT-Markov* and *BROP* components. Further, it contains an additional component called *Interest Learning Component* to find the pair-wise interest similarities between people by learning their interest. In the case of those traces that do not have social proximity (or interest) information then *Interest Learning Component* also generates synthetic interests of people. Using the sub-components of *DDT-Markov*, the `Interest-Driven Sub-Module` predicts the tighter upper bound of data dissemination time. The role of *BROP* model is same as described in `Broadcast Sub-Module` except for the additional consideration of interest similarities among people in best relay estimation.

### 3.2.4 Input Parameters

To predict the performance of data dissemination, the INDIGO framework takes several input parameters for both cases: *Contact Probability Type*, *data requirement*, *# of data sources* and *interest similarity threshold*. The *Contact Probability Type* is an input to the *Contact Probability Prediction Module* to decide about the contact patterns for prediction. For *Contact Probability Type* value as 0, INDIGO models data dissemination for static contact probabilities and if the value is 1 then it considers time-varying contact patterns. The default value of *Contact*

*Probability Type* is 0. The *data requirement* parameter is used to model the data dissemination process until the certain fraction of data is collected[5]. For broadcast strategy, we take this value as 1 while for interest-driven strategy, by default INDIGO provides several upper bounds of data dissemination time to the end user starting from Cut-off point $\alpha$ to 100% data collection with 5% (or 0.05) step size. Therefore, we get different upper bounds of data dissemination time for different data requirements i.e. $\alpha$, $\alpha + 0.05$, $\alpha + 0.10$.....1% where $\alpha$ is the Cut-off point estimated through *DDT-Markov*.

Further, *# of data sources* is used to model the multi-source data dissemination by providing the number of data sources and the *interest similarity threshold* parameter decides the threshold of interest-similarity beyond which people can exchange information under the interest-driven strategy. The default parameters for *# of data sources* is 10% of total users in the network. However for *interest similarity threshold*, INDIGO automatically estimates it though the learned interests of users by calculating the expected value of interest similarities among people.

### 3.2.5 Web Browsing History

This input is associated with a dataset to learn the interests of people using `Interest-Driven Sub-Module`. The Web Browsing History has the records of each user's visited URLs during different time periods.

### 3.2.6 Data Dissemination Performance

The *Data Dissemination Performance* stores the prediction of *Upper Bound of Data Dissemination Time* and *Best Relays* performance metrics obtained from the *Data Dissemination Prediction Module* of INDIGO framework using *DDT-Markov* and *BROP* component respectively for both cases. Please note that INDIGO focuses on providing a tighter prediction of data dissemination time.

Data dissemination time measures the time until all or some fraction of people receive information from multiple data sources and Best Relays enables the faster spread of information by finding the best nodes in the network to minimize the data dissemination time. Both metrics described above need to be predicted under different data dissemination strategies, multiple data sources and different data requirements of people.

[5]This is the maximum data requirement. It could be less in case of people leave the network or they do not meet with each other

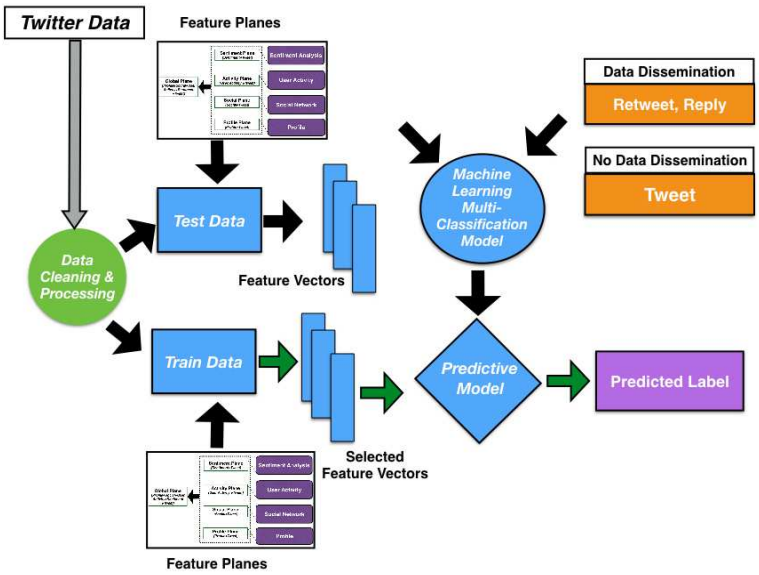



Figure 3.5. Overview of the model of INDIGO framework for Type IV case where social proximity plays the important role.

## 3.3 Overview of INDIGO for Type IV case

Figure 3.5 presents the second part of the INDIGO framework that models Type IV case of the Physical–Social proximity table using *INDIGO–OSN* part. In this case, social proximity plays the important role and predicts the performance of data dissemination in online social networks. The model predicts the performance of data dissemination using a multi-plane of features where each plane of the feature is extracted from the rich dataset of online social network based on the complexity to acquire and privacy intrusiveness. Using these feature planes the model uses a Machine Learning based model to predict the performance of data dissemination time by classifying the likelihood to retweet and reply to a certain text. More details about the complete process and model are provided in Chapter 8.

## 3.4 Conclusions

In this Chapter, I presented the overview of different parts of the INDIGO required to model and predict the performance of data dissemination process for Type II, III and IV cases of the Physical–Social proximity table. The first part *INDIGO–Physical* is dedicated to model the Type II and III cases where physi-

cal proximity is prominent and predicts the performance of data dissemination for two performance metrics i.e. *Upper Bound of Data Dissemination Time* and *Best Relays*. I presented how INDIGO addresses the Type II and Type III cases of Physical–Social proximity table by using different contact patterns and data dissemination strategies. Further, I also presented the brief overview of modeling of Type IV case using *INDIGO–OSN* part of the INDIGO framework. I summarize the key observation from this Chapter as follows:

- INDIGO provides solutions for predicting the performance of data dissemination for Type II, III and IV cases of the Physical–Social proximity table.
- For Type II and Type III cases, INDIGO provides a complete unique solution *INDIGO–Physical* that predicts the tight upper bound of data dissemination time and best relays under real-world aspects of data dissemination process.
- INDIGO utilizes five datasets from diverse environments like a conference, university and urban area called INFOCOM, PERCOM, ROLLERNET, MIT, and MACACO for Type II and III cases. Out of these 5 traces, 3 are standard traces while the other 2 (PERCOM and MACACO), I collected during my Ph.D. from conference and university environment respectively.
- To the best of my knowledge, MACACO trace is the first contact trace that collects both heterogeneous contacts and interests of people simultaneously.
- For Type IV case, INDIGO provides a Machine learning based model that predicts the performance of data dissemination using multi-plane features. Currently, INDIGO only uses the Twitter dataset.

The work done in this Chapter are presented at PerCom 2012, EWSN 2015 and Complex Networks Workshop 2016. During EWSN 2015, I also received the best poster award. In next Chapter, I will present the modeling efforts taken for Type II and III cases. More specifically, I will present the *Contact Probability Prediction Module* of INDIGO and show how it predicts the pair-wise heterogeneous contact probabilities for both static and time-varying contact patterns.

# Chapter 4

# Prediction of Heterogeneous Contact Probabilities

## 4.1 Introduction

In the previous chapter of the thesis, I gave an overview of the INDIGO data dissemination framework. In this chapter, I will focus on the *Contact Probability Prediction Module* of INDIGO that captures the contact and mobility patterns of people for Type II and Type III cases. The contact probabilities among people significantly impact the data dissemination process, therefore, the prediction of contact probabilities are an important input to the *Prediction Module* of INDIGO to enable the realistic prediction of the upper bound of data dissemination time. The *Contact Probability Prediction Module* performs this task for both static and time-varying pair-wise heterogeneous contact probabilities. For the static contact probability prediction between any pair of mobile user and the pair of mobile user and data source, I utilized Maximum Likelihood Estimation (MLE) method using either real or synthetic contact traces Picu et al. [2012]. For time-varying contact probability prediction, I use the increment learning with time and applied Machine Learning approaches by employing Gradient Boosting Machine (GBM). Figure 4.1 presents the enlarged view of our *Contact Probability Prediction Module* for both static and time-varying contact probabilities prediction.

This chapter is structured as follows. Section 4.2 presents the different challenges associated in predicting contact probabilities. In Section 4.3, I will present the method to predict heterogeneous pair-wise contact probabilities under static case from contact traces. Further, Section 4.4 presents the prediction of time-varying pair-wise contact probabilities using Machine Learning technique and also discuss the applicability of model in different contact traces. Finally, I con-

clude the Chapter with Section 4.5.

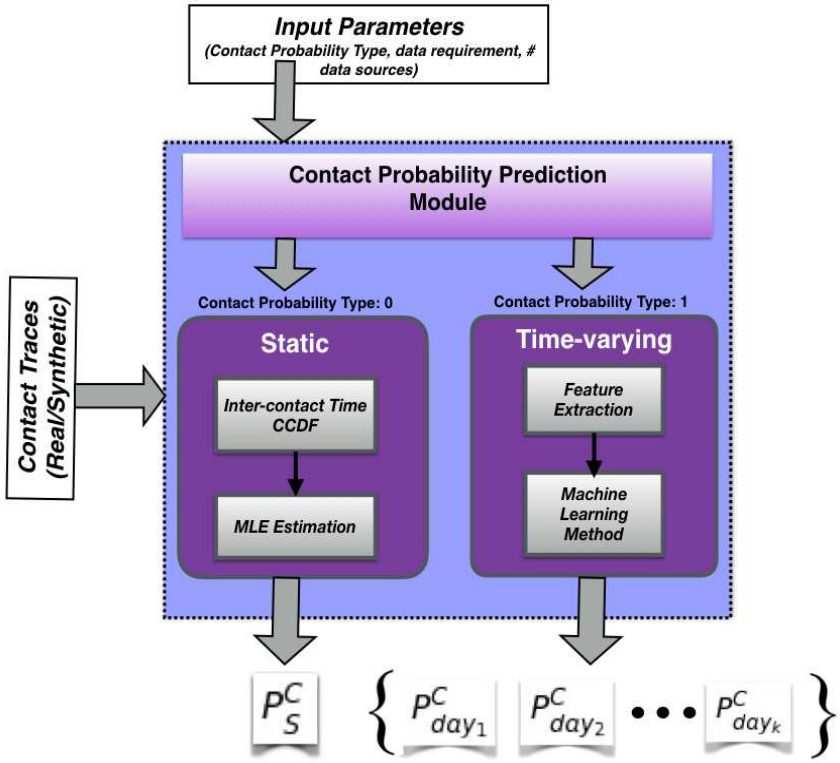



Figure 4.1. Enlarge view of Contact Probability Prediction Module for both static and time-varying contact probability prediction.

## 4.2 Prediction challenges

The ability to predict heterogeneous contact probabilities among different pair of people is quite an important problem as it enables the realistic contact patterns of people as opposed to the most of the works in literature that are either based on different mobility models Clementi et al. [2012] or homogeneous pair-wise contact rate among people Groenevelt et al. [2005]. The more precise prediction of heterogeneous contact probabilities among people will be able to provide much better and tighter prediction of the upper bound of data dissemination time. The prediction of static heterogeneous contact probabilities among people has drawn some attention in the literature and some of the work has been in this direction by utilizing the distribution of inter-contact times Conan et al. [2006] Picu et al. [2012] Boldrini et al. [2014]. In this thesis, I also adopt the similar approach

by utilizing the inter-contact time distribution and MLE method as this method has been outperformed as compared to other methods. Further, the time-varying heterogeneous contact probability prediction is still not addressed in literature even though it represents more realistic aspects of human mobility. Some of the challenges for the prediction of time-varying contact patterns is how do we find the optimum time slots for contact probability estimation; how do we learn the patterns of contacts among people and take into account the contact patterns among people on different days of the week or hours of a day. Since time-varying contact probability is not well addressed in the literature, therefore, the another challenge for such prediction is the importance of different parameters (or predictors) in predicting time-varying heterogeneous pair-wise contact probabilities. In this Chapter, I address these challenges and discuss different approaches utilized in predicting time-varying contact probabilities.

## 4.3 Pair-wise static heterogeneous contact probabilities prediction

In this Section, I present an approach to predict static pair-wise heterogeneous contact probabilities using `Static Component` of the *Contact Probability Prediction Module* shown in Figure 4.1. While considering static pair-wise contact probabilities, I assume that each pair of people will exhibit same contact patterns during different time periods. INDIGO provides a way to predict static contact probabilities and enables the prediction of the upper bound of data dissemination time under heterogeneous static contact patterns. The static contact patterns are useful for the contact traces with small duration of data. For example, in case of conferences (INFOCOM Scott et al. [2009], PERCOM SCAMPI [2012]) or urban area (ROLLERNET Benbadis and Leguay [2009]) traces with small duration, I can consider static contact patterns among people because variation over time is negligible.

To predict the static pair-wise heterogeneous contact probabilities among people and different data sources, `Static Component` first finds the distribution of inter-contact time for different contact traces followed by employing the MLE method. To understand the contact probability prediction using MLE, I first define contact and inter-contact time.

*Definition 1 (Contact): Under heterogeneous mobility, we define contact between any two nodes i and j when they come in the communication range of each other and able to exchange information with each other.*

*Definition 2 (Inter-contact Time): The time interval $\Delta T_{ij}$ between two successive contacts of a node pair $i, j$ is defined as inter-contact time.*

I define $\Delta T_{ij}$ as follows:

$$\Delta T_{ij} = T_{ij}^{m} - T_{ij}^{n} \tag{4.1}$$

Where $T_{ij}^{m}$ and $T_{ij}^{n}$ present the starting time for the contact of $i, j$ node pair at two time slots $m$ and $n$. Figure 4.2 presents the graphical representation of inter-contact time between node pair $i, j$ where the pair meets at time slot $m$ and $n$ and the contact ends at time slot $m+k$ and $n+p$.

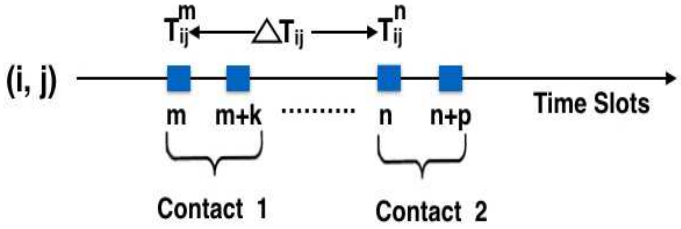



Figure 4.2. Graphical representation of Inter-contact time between a node pair $i, j$ when they contact at time slot at time $m$ and $n$.

### 4.3.1 Distribution of inter-contact time

To employ Maximum Likelihood Estimation (MLE), I first need to understand the distribution of inter-contact times between each pair. Inter-contact times are generally very important to data dissemination time analyses as they determine the delay of data exchange in a network. Thus, they have been extensively studied, both via experimental traces and by analyzing simple mobility models. These studies have concluded that three main distribution types appear in inter-contact time: the exponential distribution, the power law, and the power law with exponential cutoff.

The exponential distribution is considered because of its memorylessness property, the simplest of the three. As such, it is also the one most often used both in analyses and in algorithm design and simulation. This distribution is motivated via the analysis and simulation of simple mobility models, such as random waypoint or random direction Groenevelt et al. [2005]. However, from experimental traces, the inter-contact times aggregate over all node pairs were found to be power-law distributed Scott et al. [2006a]. This caused a certain dismay in the research community as it meant that the average delays in routing and data dissemination algorithms could potentially be infinite. Nevertheless, a wave of new analytical studies Karagiannis et al. [2007] Cai and Eun [2009] combined with closer looks at experimental traces have dismissed these concerns. Cai et al. showed in Cai and Eun [2009] that, under relatively generic conditions, the pair-wise inter-contact times have a probability distribution, which is a mixture of a power law and an exponential (i.e., power-law head and exponential tail). This has also been confirmed in real-world traces by Karagiannis et al. Karagiannis et al. [2007], who analyzed the empirical complementary cumulative distribution functions (CCDFs) of the inter-contacts and defined as follows Picu et al. [2012]:

$$F_{ij}(x) = \begin{cases} C_{ij}.x^{-\alpha_{ij}}e^{-\beta_{ij}x} & for \quad x \geq t_0^{ij} \\ 1 & for \quad 0 < x < t_0^{ij} \end{cases}$$

Where $t_0^{ij}$ is the minimum inter-contact time between the $i, j$ node pair and $C_{ij} = (t_0^{ij})^{\alpha_{ij}e^{\beta_{ij}t_0^{ij}}}$ is a positive normalization constant. The above function is the combination of a power law distribution and an exponential distribution. The two inter-contact time distribution parameters $\alpha$ and $\beta$ represent power law exponent of inter-contact times of the nodes pair $i, j$ and contact rate of node pair $i, j$. To estimate $\alpha$ and $\beta$ parameters for each node pair $i, j$, I employ Maximum Likelihood Estimation (MLE) by utilizing the CCDF function $F_{ij}(x)$.

### 4.3.2 Maximum Likelihood Estimation method

The Maximum Likelihood Estimation (MLE) method was introduced in 1921 by Sir Ronald Fisher and chooses the estimate of the parameter which "makes the observed data as likely as possible" Scholz [1985]. It is a method of estimating the parameters of a statistical model given observations, by finding the parameter values that maximize the likelihood of making the observations given the

parameters. In general, for a fixed set of data and underlying statistical model, the method of maximum likelihood selects the set of values of the model parameters that maximize the likelihood function. Intuitively, this maximizes the "agreement" of the selected model with the observed data, and for discrete random variables it indeed maximizes the probability of the observed data under the resulting distribution.

*Definition 3: If the sample data is denoted by $x$, the parameter by $\theta$ and the probability density function by $f(x;\theta)$ then the maximum likelihood estimate of $\theta$ is that value of $\theta$, $\hat{\theta}$ which maximizes $f(x;\theta)$.*

Maximum likelihood estimates are obtained by maximizing the likelihood using calculus. Most often we have a random sample of size $n$ from a population with density function $f(x;\theta)$. In this case, we have maximum likelihood estimate $L(x;\theta)$ as :

$$L(x;\theta) = \prod_{k=1}^{n} f(x_k;\theta) \tag{4.2}$$

Since the maximum of a function occurs at the same value as the maximum of the natural logarithm of the function and it will also be easier to maximize with respect to $\theta$

$$\sum_{k=1}^{n} ln[f(x_k;\theta)] \tag{4.3}$$

Finally, I solve the following equation to estimate $\hat{\theta}$ parameter which is called the maximum likelihood.

$$\sum_{k=1}^{n} \frac{\partial ln[f(x_k;\theta)]}{\partial \theta} = 0 \tag{4.4}$$

### 4.3.3 Pair-wise contact probability prediction using Maximum Likelihood Estimation

For the contact traces, we estimate contact probabilities utilizing the MLE method described in Section 4.3.2. In this section, we present the methodology to employ MLE to predict the contact probability utilizing the inter-contact time CCDF function $F_{ij}(x)$ derived in Section 4.3.1.

*Definition 4 (Pair-wise Contact Probability): The pair-wise contact probability $p_{ij}^{c}$ between any node-pair, $i$ and $j$ is defined as the likelihood of meeting the given pair with respect to the contacts with other nodes.*

The contact probability is calculated using the contact rate $\beta_{ij}$ between the node pair and defined as follows:

$$p_{ij}^{c} = \frac{\beta_{ij}}{\sum_{1 \leq = i < j \leq N} \beta_{ij}} \tag{4.5}$$

Where $N$ is the total number of nodes in the network. The contact rate $\beta_{ij}$ is estimated using MLE method as follows:

- Calculate the Probability Density Function (PDF) of inter-contact time distribution for the node pair $i$ and $j$ using CCDF function $F_{ij}(x)$ of inter-contact time derived in Section 4.3.1.

$$f_{ij}(x) = \frac{\partial (1 - F_{ij}(x))}{\partial x} \tag{4.6}$$

- Utilize Maximum Likelihood Estimation to estimate $\alpha_{ij}$ and $\beta_{ij}$.

$$L_{ij}* = ln(L(x; \alpha_{ij}, \beta_{ij})) = \sum_{k=1}^{N} ln[f_{ij}(x_k; \alpha_{ij}, \beta_{ij})] \tag{4.7}$$

- Find contact rate parameter $\beta_{ij}$.

$$\frac{\partial L_{ij}*}{\partial \alpha_{ij}} = 0, \quad \frac{\partial L_{ij}*}{\partial \beta_{ij}} = 0 \tag{4.8}$$

- Finally, estimate the contact probability of pair $(i, j)$ in $N$ node network using Equation 4.5.

Likewise, I estimate the heterogeneous contact probabilities between each pair of nodes and construct the contact probability matrix $P^C$ that contains *pair-wise heterogeneous contact probability* between each pair of nodes in the network $V$.

$$P_S^C = \{p_{ij}^c\}\ i, j \in V \tag{4.9}$$

Figure 4.3 presents the sample static heterogeneous contact probabilities predicted from all contact traces considered in this thesis using MLE.

## 4.4 Pair-wise time-varying heterogeneous contact probabilities prediction

In this Section, I will focus on predicting time-varying pair-wise contact probabilities utilizing the `Time-varying Component` of the *Contact Probability Prediction Module* shown in Figure 4.1. To the best of my knowledge, the work done in literature only consider the static heterogeneous contact probabilities among people for data dissemination, however, as discussed in previous Chapter 1 2, the assumption of static contact probabilities do not hold in reality as people do not meet with each other with same likelihood. For example, people do meet their colleagues more on weekdays as compared to weekends therefore, one cannot assume the static contact probabilities over time. Further, the assumption of static contact probabilities does not provide a realistic prediction of data dissemination process as it cannot capture varying mobility patterns of people over time. In this thesis, I address the problem of time-varying contact probabilities and provide a solution to learn contact probabilities of people over time. My approach is based on Machine Learning methods that meet the challenges described in Section 4.4.1 by learning the contact patterns of people over time.

### 4.4.1 A Machine Learning approach

For the time-varying contact patterns case, one of the important challenges is to automatically find the optimum time slot for different people. One of the possible solutions is to divide the total contact traces in a fixed time slot followed by the prediction of contact probabilities using MLE. However, this approach will not capture the true behavior of human mobility patterns and will always rely on the availability of complete contact traces which is not always feasible to collect. The work done in Gao et al. [2013] has shown that people show transient

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| 0 | 0.000000 | 0.000116 | 0.001586 | 0.000081 | 0.000012 | 0.000000 | 0.00000 |
| 1 | 0.000116 | 0.000000 | 0.009943 | 0.004526 | 0.016669 | 0.004051 | 0.00000 |
| 2 | 0.001586 | 0.009943 | 0.000000 | 0.028487 | 0.032550 | 0.004688 | 0.00000 |
| 3 | 0.000081 | 0.004526 | 0.028487 | 0.000000 | 0.004237 | 0.026056 | 0.00000 |
| 4 | 0.000012 | 0.016669 | 0.032550 | 0.004237 | 0.000000 | 0.000035 | 0.00000 |
| 5 | 0.000000 | 0.004051 | 0.004688 | 0.026056 | 0.000035 | 0.000000 | 0.00000 |
| 6 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.00000 |
| 7 | 0.000035 | 0.004225 | 0.030744 | 0.012501 | 0.012478 | 0.049878 | 0.00000 |
| 8 | 0.004179 | 0.015395 | 0.019817 | 0.000324 | 0.065737 | 0.000023 | 0.00000 |
| 9 | 0.000000 | 0.016935 | 0.029309 | 0.008624 | 0.006910 | 0.000059 | 0.00000 |
| 10 | 0.100340 | 0.002929 | 0.000000 | 0.113670 | 0.000012 | 0.011217 | 0.00000 |
| 11 | 0.000000 | 0.000000 | 0.000000 | 0.000046 | 0.000000 | 0.000000 | 0.00000 |
| 12 | 0.001435 | 0.002651 | 0.000220 | 0.000035 | 0.000000 | 0.000012 | 0.00000 |
| 13 | 0.046684 | 0.075715 | 0.039819 | 0.008415 | 0.012432 | 0.004248 | 0.00000 |
| 14 | 0.002766 | 0.004410 | 0.081433 | 0.019863 | 0.011668 | 0.023915 | 0.00000 |
| 15 | 0.033731 | 0.007038 | 0.026948 | 0.022792 | 0.038963 | 0.000023 | 0.00000 |
| 16 | 0.025570 | 0.000197 | 0.018752 | 0.050446 | 0.020894 | 0.001563 | 0.00000 |
| 17 | 0.004017 | 0.002801 | 0.000139 | 0.012640 | 0.007084 | 0.115050 | 0.53161 |
| 18 | 0.071756 | 0.000023 | 0.000000 | 0.000035 | 0.013879 | 0.000000 | 0.00000 |
| 19 | 0.088378 | 0.004283 | 0.010071 | 0.004260 | 0.001331 | 0.000012 | 0.00000 |
| 20 | 0.001621 | 0.032643 | 0.018532 | 0.038349 | 0.011055 | 0.000023 | 0.00000 |
| 21 | 0.035756 | 0.000035 | 0.017386 | 0.023637 | 0.004526 | 0.013983 | 0.00000 |
| 22 | 0.004190 | 0.055724 | 0.015951 | 0.046059 | 0.001551 | 0.101990 | 0.00000 |
| 23 | 0.060146 | 0.050284 | 0.023475 | 0.012756 | 0.024771 | 0.006876 | 0.00000 |
| 24 | 0.035930 | 0.004283 | 0.087406 | 0.021333 | 0.014145 | 0.020002 | 0.00000 |
| 25 | 0.000023 | 0.004271 | 0.039623 | 0.008612 | 0.002801 | 0.021125 | 0.00000 |

(a) Conference Trace–INFOCOM

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| 0 | 0.000000 | 0.001085 | 0.027587 | 0.000018 | 0.026554 | 0.057572 | 0.005391 |
| 1 | 0.001085 | 0.000000 | 0.006459 | 0.004236 | 0.002223 | 0.025364 | 0.008542 |
| 2 | 0.027552 | 0.006459 | 0.000000 | 0.000000 | 0.052864 | 0.200310 | 0.018957 |
| 3 | 0.000018 | 0.004236 | 0.000000 | 0.000000 | 0.131480 | 0.129660 | 0.017785 |
| 4 | 0.026554 | 0.002223 | 0.052864 | 0.131480 | 0.000000 | 0.373440 | 0.188630 |
| 5 | 0.057572 | 0.025364 | 0.200310 | 0.129660 | 0.373440 | 0.000000 | 0.173860 |
| 6 | 0.005391 | 0.008542 | 0.018957 | 0.017785 | 0.188630 | 0.173860 | 0.000000 |
| 7 | 0.013916 | 0.010275 | 0.045074 | 0.038597 | 0.158240 | 0.167170 | 0.134370 |
| 8 | 0.011746 | 0.002188 | 0.057047 | 0.029355 | 0.071384 | 0.209760 | 0.111750 |
| 9 | 0.013934 | 0.010800 | 0.011623 | 0.000018 | 0.040873 | 0.058973 | 0.011501 |
| 10 | 0.011658 | 0.000998 | 0.000000 | 0.000000 | 0.035044 | 0.030668 | 0.000053 |
| 11 | 0.028235 | 0.000000 | 0.041818 | 0.000000 | 0.070683 | 0.007610 | 0.012936 |
| 12 | 0.001085 | 0.001068 | 0.000000 | 0.002188 | 0.038930 | 0.035569 | 0.005129 |
| 13 | 0.005461 | 0.016594 | 0.013391 | 0.006232 | 0.030493 | 0.019623 | 0.035307 |
| 14 | 0.023491 | 0.000000 | 0.038808 | 0.019220 | 0.105760 | 0.062439 | 0.135200 |
| 15 | 0.008402 | 0.000000 | 0.014406 | 0.000000 | 0.010468 | 0.018485 | 0.012516 |
| 16 | 0.034571 | 0.009382 | 0.002416 | 0.000000 | 0.037232 | 0.060216 | 0.004289 |
| 17 | 0.057677 | 0.018887 | 0.032278 | 0.000123 | 0.050676 | 0.088240 | 0.032996 |
| 18 | 0.069563 | 0.018187 | 0.082131 | 0.012638 | 0.120520 | 0.182360 | 0.079663 |
| 19 | 0.047227 | 0.014634 | 0.026064 | 0.000018 | 0.061056 | 0.103400 | 0.005409 |
| 20 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.006442 | 0.000000 | 0.006232 |
| 21 | 0.002993 | 0.000000 | 0.000000 | 0.000000 | 0.003168 | 0.000018 | 0.001015 |

(b) University Trace–PERCOM

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| 0 | 0.000000 | 0.000000 | 0.000028 | 0.000094 | 0.000053 | 0.000041 | 0.000000 |
| 1 | 0.000000 | 0.000000 | 0.000003 | 0.000005 | 0.000002 | 0.000000 | 0.000043 |
| 2 | 0.000028 | 0.000003 | 0.000000 | 0.000017 | 0.000129 | 0.000002 | 0.000000 |
| 3 | 0.000094 | 0.000005 | 0.000017 | 0.000000 | 0.000301 | 0.000279 | 0.000002 |
| 4 | 0.000053 | 0.000002 | 0.000129 | 0.000301 | 0.000000 | 0.000235 | 0.000005 |
| 5 | 0.000041 | 0.000000 | 0.000002 | 0.000279 | 0.000235 | 0.000000 | 0.000020 |
| 6 | 0.000000 | 0.000043 | 0.000000 | 0.000002 | 0.000005 | 0.000020 | 0.000000 |
| 7 | 0.000225 | 0.000245 | 0.000012 | 0.000126 | 0.000015 | 0.000025 | 0.000003 |
| 8 | 0.000000 | 0.000002 | 0.000003 | 0.000002 | 0.000002 | 0.000000 | 0.000000 |
| 9 | 0.000093 | 0.000026 | 0.000003 | 0.000015 | 0.000005 | 0.000008 | 0.000003 |
| 10 | 0.000000 | 0.000281 | 0.000007 | 0.000005 | 0.000015 | 0.000000 | 0.000051 |
| 11 | 0.000060 | 0.000005 | 0.000022 | 0.000008 | 0.000003 | 0.000000 | 0.000283 |
| 12 | 0.000002 | 0.000031 | 0.000000 | 0.000000 | 0.000002 | 0.000000 | 0.000000 |
| 13 | 0.000099 | 0.000002 | 0.000028 | 0.000374 | 0.000187 | 0.000223 | 0.000040 |
| 14 | 0.000053 | 0.000341 | 0.000000 | 0.000002 | 0.000000 | 0.000017 | 0.000007 |
| 15 | 0.000000 | 0.000003 | 0.000000 | 0.000000 | 0.000002 | 0.000000 | 0.000048 |
| 16 | 0.000000 | 0.000000 | 0.000000 | 0.000003 | 0.000000 | 0.000000 | 0.000000 |
| 17 | 0.000000 | 0.000000 | 0.000000 | 0.000005 | 0.000000 | 0.000000 | 0.000000 |
| 18 | 0.000071 | 0.000002 | 0.000060 | 0.000096 | 0.000056 | 0.000041 | 0.000008 |
| 19 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| 20 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| 21 | 0.000002 | 0.000028 | 0.000000 | 0.000003 | 0.000005 | 0.000005 | 0.000293 |
| 22 | 0.000013 | 0.000002 | 0.000005 | 0.000000 | 0.000007 | 0.000002 | 0.000013 |
| 23 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| 24 | 0.000000 | 0.000002 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| 25 | 0.000002 | 0.000018 | 0.000000 | 0.000002 | 0.000003 | 0.000002 | 0.000007 |
| 26 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| 27 | 0.000013 | 0.000000 | 0.000000 | 0.000000 | 0.000002 | 0.000000 | 0.000000 |
| 28 | 0.000002 | 0.000002 | 0.000000 | 0.000038 | 0.000041 | 0.000116 | 0.000000 |
| 29 | 0.000000 | 0.000002 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| 30 | 0.000000 | 0.000023 | 0.000002 | 0.000002 | 0.000000 | 0.000007 | 0.000008 |

(c) University Trace–MIT

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| 0 | 0.000000 | 0.052125 | 0.043204 | 0.031676 | 0.032879 | 0.026263 | 0.036287 |
| 1 | 0.052125 | 0.000000 | 0.013532 | 0.003408 | 0.042402 | 0.007719 | 0.005313 |
| 2 | 0.043204 | 0.013532 | 0.000000 | 0.019948 | 0.030473 | 0.035285 | 0.019948 |
| 3 | 0.031676 | 0.003408 | 0.019948 | 0.000000 | 0.006415 | 0.009723 | 0.089415 |
| 4 | 0.032879 | 0.042402 | 0.030473 | 0.006415 | 0.000000 | 0.036488 | 0.019046 |
| 5 | 0.026263 | 0.007719 | 0.035285 | 0.009723 | 0.036488 | 0.000000 | 0.021953 |
| 6 | 0.036287 | 0.005313 | 0.019948 | 0.089415 | 0.019046 | 0.021953 | 0.000000 |
| 7 | 0.017041 | 0.043905 | 0.025762 | 0.007217 | 0.008921 | 0.013031 | 0.011026 |
| 8 | 0.024258 | 0.039996 | 0.027466 | 0.030273 | 0.010826 | 0.006816 | 0.040096 |
| 9 | 0.018043 | 0.027767 | 0.052225 | 0.000000 | 0.012731 | 0.009723 | 0.000000 |
| 10 | 0.040998 | 0.190557 | 0.010926 | 0.001103 | 0.017342 | 0.001403 | 0.011828 |
| 11 | 0.021852 | 0.020349 | 0.016740 | 0.023557 | 0.022955 | 0.025461 | 0.065457 |
| 12 | 0.004611 | 0.016840 | 0.015738 | 0.063252 | 0.029370 | 0.025461 | 0.067261 |
| 13 | 0.008520 | 0.004411 | 0.010926 | 0.028669 | 0.010726 | 0.011227 | 0.044306 |
| 14 | 0.007017 | 0.049719 | 0.044006 | 0.021752 | 0.012129 | 0.021050 | 0.013432 |
| 15 | 0.017442 | 0.023055 | 0.004411 | 0.020048 | 0.019848 | 0.015136 | 0.036488 |
| 16 | 0.062149 | 0.243384 | 0.013332 | 0.000702 | 0.027666 | 0.006315 | 0.011427 |
| 17 | 0.008019 | 0.004812 | 0.004110 | 0.097835 | 0.006716 | 0.012129 | 0.019346 |
| 18 | 0.014735 | 0.005513 | 0.015437 | 0.031275 | 0.012029 | 0.010525 | 0.019046 |
| 19 | 0.004812 | 0.001303 | 0.005714 | 0.021752 | 0.021752 | 0.026163 | 0.018745 |
| 20 | 0.273757 | 0.039896 | 0.037490 | 0.015838 | 0.025962 | 0.030373 | 0.024459 |

(d) Crowded Urban Trace–ROLLERNET

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| 0 | 0.000000 | 0.000007 | 0.000222 | 0.000000 | 0.000053 | 0.000000 | 0.000000 |
| 1 | 0.000007 | 0.000000 | 0.003158 | 0.000000 | 0.004355 | 0.000000 | 0.000000 |
| 2 | 0.000222 | 0.003158 | 0.000000 | 0.000013 | 0.003960 | 0.000000 | 0.000002 |
| 3 | 0.000000 | 0.000000 | 0.000013 | 0.000000 | 0.000030 | 0.000008 | 0.000050 |
| 4 | 0.000053 | 0.004355 | 0.003960 | 0.000030 | 0.000000 | 0.000000 | 0.000013 |
| 5 | 0.000000 | 0.000000 | 0.000000 | 0.000008 | 0.000000 | 0.000000 | 0.000003 |
| 6 | 0.000000 | 0.000000 | 0.000002 | 0.000050 | 0.000013 | 0.000003 | 0.000000 |

(e) Our Trace–MACACO

Figure 4.3. A sample of heterogeneous pair-wise static contact probabilities estimated for the diverse environment.

contact patterns in different time slots in a day which does not necessarily be symmetric and can be different for different days. Further, recent works have shown that over a longer duration, people do exhibit regularity in their mobility patterns Gonzalez et al. [2008] Hsu et al. [2009] and that, in context of social and location-based networks, Machine Learning approaches can be used for link prediction Liben-Nowell and Kleinberg [2007] Scellato et al. [2011].

Therefore, to predict time-varying contact patterns, I also adopt Machine Learning based approach. In this analysis, I consider the MIT and MACACO trace as they have the contact patterns of people with longer duration. In context to INFOCOM, PERCOM and ROLLERNET traces, where the contact patterns are random, I rely on the static contact probabilities. Machine Learning approach enables the automatic learning of heterogeneous contact patterns among people and also provide us the most important predictors required to predict contact probabilities under time-varying patterns.

### 4.4.2 Time-varying contact probability prediction model

Figure 4.4 presents the overall procedure to predict the time-varying contact probabilities of `Time-varying Component` of Figure 4.1. After taking the contact traces as input, the model first process the data for each week and then create daily features by incremental updating the features of the same day-of-the-week of the previous week. For example, if we have 2 weeks of data then, I create features for the Monday of the first week by considering only the data of Monday while for the Monday of the second week, I upgrade the features of the Monday of week 1 with the ones of the current Monday (of week 2). Likewise, I create daily features for all contact traces and learn the daily contact patterns of people in this manner [1]

#### Feature development

Features define the attribute or properties of the data set. I create features for each pair of people. For each pair, the features considered in my model are related to the *contact pattern statistics of each person in the pair with others*, *contact pattern statistics between the given pair* and *the inter-contact time patterns statistics* and their *distribution parameters*. In addition to these features, I also have

[1]Please note that I also tried to create hourly features for each day of each week, however, the results were not promising thus, shows that such granular contact patterns prediction is quite difficult to achieve due to sparse contact data. In the case of granular contact datasets, the model can be further tested hourly basis or another appropriate time slots.

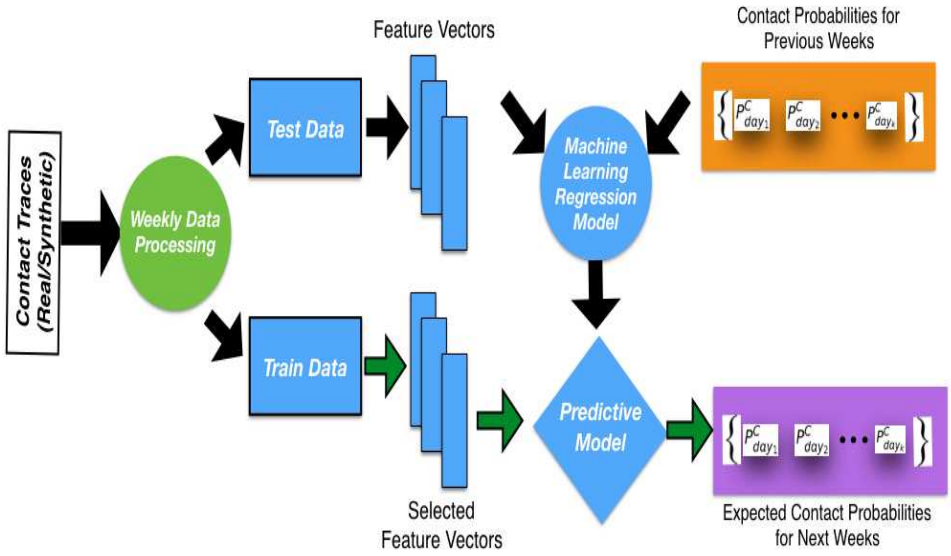



Figure 4.4. The process to predict time-varying contact probabilities using Machine Learning Approach.

day of the week as a feature. These features capture the overall contact pattern of a person with all other people as well as his/her contact pattern with the individual person of the pair. The contact related features inspect the community formation and ties of a person with others in the network while the inter-contact time features to capture the mobility of the person i.e. how often he/she is interested in meeting with other people at different times. Table 4.1 presents all features extracted from the contact traces extracted for a node pair $i, j$ for a certain day of the week. The total number of features developed from the datasets are 50.

### XGBoost: A Gradient Boosting Machine method

The contact probability between any pair of people at a given day is a continuous variable that lies between 0 to 1. Therefore, contact probability prediction can be seen as a regression problem where the aim of my model is to minimize the distance between the predicted and the observed value (error). To do this, I analyzed several Machine Learning methods like Linear Regression Weisberg [2005], Random Forest Breiman [2001], Support Vector Machines (SVMs) Suykens and Vandewalle [1999] and Gradient Boosting Machine Friedman [2001]. Out of all these methods, Gradient Boosting Method (GBM) outperforms for different samples of contact traces. In Linear Regression method, the error was high as

Table 4.1. Features Extracted from Contact Traces

| | |
|---|---|
| Contact Features | Total # Contacts of $i$ and $j$ with all people, Descriptive statistics (mean, median, std, quantiles) of # Contacts of $i$ and $j$, # people contacted for $i$ and $j$, Contact Likelihood Hourly of $i$ and $j$ |
| Inter Contact Time (ICT) Features | Descriptive statistics (mean, median, std, quantiles) of ICT of $i$ and $j$ with all people, parameters of ICT distribution for $i$ and $j$ with all people |
| Node Pair Features | Total # Contacts between $i$ and $j$, Descriptive statistics (mean, median, std, quantiles) of contact time between $i$ and $j$, Descriptive statistics (mean, median, std, quantiles) of ICT between $i$ and $j$, parameters of ICT distribution between $i$ and $j$ |
| Time Features | Week Number, Day Number |

there the Contact probability and predictor parameter are not linearly related while Random Forest was not able to provide better prediction method since it's a bagging method that mainly relies on the voting or averaging method. In the case of GBM, its nature of boosting helps to correct the error in each sequence as it works on sequential methods. Therefore, to predict time-varying contact probabilities among people, I finally considered Gradient Boosting Machine method. More specifically, I use the XGBoost or "Extreme Gradient Boosting" method.

Gradient Boosting is a supervised machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It is a sequential technique which works on the principle of ensemble and combines a set of weak learners and delivers improved prediction accuracy. At any instant $t$, the model outcomes are weighed based on the outcomes of previous instant $t-1$. The outcomes predicted correctly are given a lower weight and the ones miss-classified are weighted higher. Gradient Boosting algorithms play a crucial role in dealing with bias and variance trade-off.

Out of these Gradient Boosting algorithms, I choose XGBoost as it uses a more regularized model formalization to control over-fitting hence gives better performance. Further, XGBoost also provides much faster computation for boosted tree algorithms and makes it more suitable for large contact traces. I highlight some of the advantages of XGBoost over traditional GBM as follows:

- **Regularization**: Standard GBM implementation has no regularization like XGBoost, therefore it also helps to reduce over-fitting. In fact, XGBoost is also known as "regularized boosting" technique.

- **Parallel Processing**: XGBoost implements parallel processing and is amazingly faster as compared to GBM.

- **High Flexibility**: XGBoost allows users to define custom optimization objectives and evaluation criteria. This adds a whole new dimension to the model and there is no limit to what we can do.

- **Handling Missing Values**: XGBoost has an inbuilt routine to handle missing values. XGBoost tries different things as it encounters a missing value on each node and learns which path to take for missing values in future.

- **Tree Pruning**: A GBM would stop splitting a node when it encounters a negative loss in the split. Thus it is more of a greedy algorithm. XGBoost, on the other hand, makes splits upto the max_depth specified and then start pruning the tree backward and remove splits beyond which there is no positive gain. Another advantage is that sometimes a split of negative loss say -2 may be followed by a split of positive loss +10. GBM would stop as it encounters -2. But XGBoost will go deeper and it will see a combined effect of +8 of the split and keep both.

- **Built-in Cross-Validation**: XGBoost allows a user to run a cross-validation at each iteration of the boosting process and thus it is easy to get the exact optimum number of boosting iterations in a single run. This is unlike GBM where we have to run a grid-search and only a limited values can be tested.

My implementation of Gradient Boosting Method is based on the Python library XGBoost[2]. I use the *XGBRegressor* function of XGBoost to predict the contact probabilities among people. Further, I tried a set of parameter combinations to prevent over-fitting using two parameters, *eta* that determines the learning rate and the number of rounds i.e. *n_estimators*. I experimentally set optimum *eta* and *n_estimators* for each trace. I also apply 10-fold cross-validation to select an appropriate number of rounds based on the mean error rate. From the contact traces, I predict the contact probabilities incrementally over time. As explained above, my model first learns the contact patterns among people using the first week of the data and predicts contact probabilities for the 2nd week. Similarly

[2]xgboost.readthedocs.io/en/latest/python/python intro.html

to predict contact probabilities for 3rd week, it learns contact patterns from both week 1 and week 2 data. Unlike existing XGBoost algorithm that does not work on time-series testing, I modified the algorithm to enable time-series prediction. For all traces, the model predicts the daily time-varying pair-wise contact probabilities for different weeks. I present the set of time-varying contact probabilities $P^C_{TV}$ predicted for a certain week $w$ having 7 days starting from Monday (1) to Sunday (7) as follows:

$$P^C_{TV} = \{P^C_{day_1}, P^C_{day_2}.....P^C_{day_k}\}\ i,j \in V\ k \in [1,7] \tag{4.10}$$

Where,

$$P^C_{day_k} = \{p_{ij}^{c_{day_k}}\}\ i,j \in V \tag{4.11}$$

$P^C_{day_k}$ represents the contact probability matrix between each pair of users for a given day $day_k$ of week $w$ and $p_{ij}^{c_{day_k}}$ presents the predicted contact probability value between the pair $i,j$ for the $k^{th}$ day.

### Feature selection

Another important aspect of prediction is to find the most important features that define the target variable (in my case it's contact probability between any pair of people). If we have too many features then it might lead to over fitting and does not provide accurate results for testing data. Therefore, to overcome this problem I applied different feature selection methods on the training data and find the set of most important features that contributes most to learning contact patterns. Out of these important features, I used top-k approach and gave a different set of features to the prediction model. From our results, we see that only 20 features (or predictors) are enough to predict time-varying pair-wise contact probabilities among people. For feature selection, I used one of the most widely used Recursive Feature Elimination (RFE) technique on the linear regression model. Given an external estimator that assigns weights to features (e.g., the coefficients of a linear model), the goal of RFE is to select features by recursively considering smaller and smaller sets of features. First, the estimator is trained on the initial set of features and weights are assigned to each one of them. Then, features whose absolute weights are the smallest are pruned from the current set features. That procedure is recursively repeated on the pruned set until the desired number of features to select is eventually reached. I present the top or most important features used in prediction for both MIT and MACACO contact traces in next section.

### 4.4.3 Results

In this Section, I will present the predicted time-varying contact probabilities obtained using the `Time-varying Component` of the *Contact Probability Prediction Module* of INDIGO. As explained in Section 4.4.1, I present predicted contact probabilities only for MIT and MACACO traces. For both traces, I validate the time-varying prediction model for one month contact data. For MIT trace, I divide the month data in 4 weeks and process each week data to create daily features (described in Section 4.4.2). I call these weeks as MIT-W1, MIT-W2, MIT-W3 and MIT-W4 respectively. For our own collected MACACO traces from different countries and groups, I also consider one month contact data where a maximum number of users make regular and intense use of the wireless network. The most active month for France and Brazil group were May 2015 and October 2015 respectively. For France trace, there were only three weeks (France-W1, France-W2, and France-W3) where we have enough contacts of people, therefore, I present prediction for France-W2, and France-W3. Similarly, for Brazil trace, since we had contacts only for 2 weeks, therefore, I present our contact probabilities prediction results for Brazil-W2 [3]. Similarly for MACACO traces, I also process the each week data and create daily features.

After creating daily features, I train the prediction model with the first week of data for both traces and then test the model by predicting the contact probabilities between different pairs of people for the next week. Likewise, I keep on predicting contact probabilities for all weeks of both traces. In the case of MIT, I train the model with MIT-W1 data and subsequently predict the contact probabilities for MIT-W2, MIT-W3, and MIT-W4 respectively. For MACACO France and Brazil traces, I test my prediction model for France-W2, France-W3, and Brazil-W2 respectively. Further, I calculate the mean *Mean Absolute Percentage Error (MAPE* obtained for each day between the observed and predicted value for all pair of users. MAPE is a non-scaled error metric that is used as a figure of merit to identify whether a perdition method is performing well or not. The lower the MAPE, the better the performance of the model. This measure is easy to understand because it provides the error in terms of percentages. I used this metric as I wanted to see the absolute error in predicting the contact probabilities that do not get impacted by the cancellation of positive and negative errors. For the prediction results I calculated the Mean Absolute Percentage Accuracy (MAPA) for each day of the week and defined as follows:

[3] Please note that throughout the thesis, I considered the same month for the validation of our different results obtained from INDIGO.

$$MAPA_{day_k} = 100 - \left(\frac{1}{n}\sum \frac{|ActualCP - PredictedCP|}{|ActualCP|}\right) * 100 \tag{4.12}$$

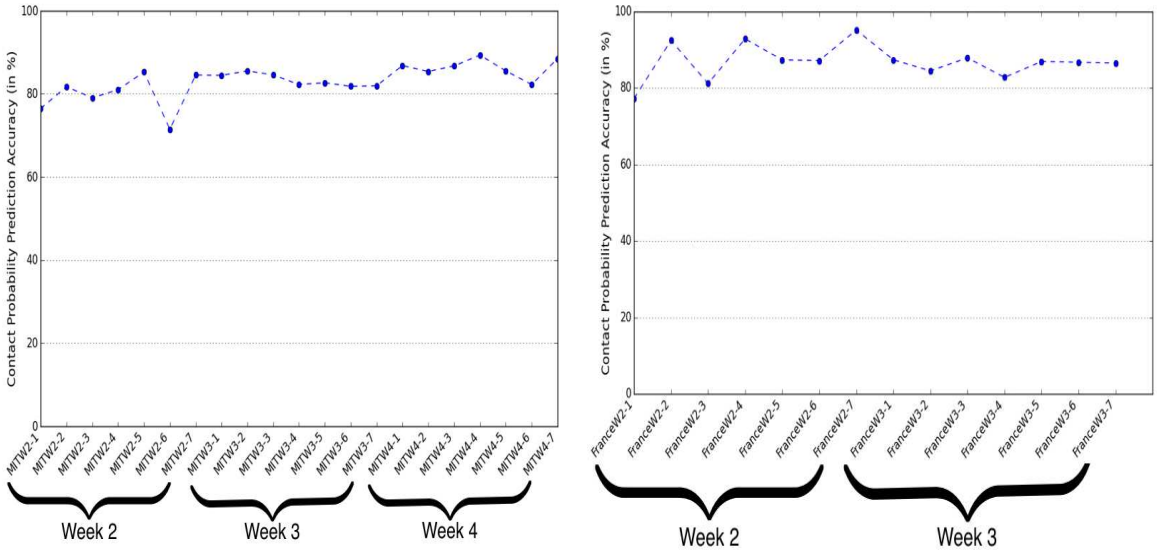
Figure 4.5 presents the contact Mean Percentage Accuracy obtained from the prediction model for both MIT and MACACO traces. From Figure 4.5 we observe that the proposed model is able to predict time-varying contact probabilities between 80-90% accuracy for MIT and MACACO-France traces. Further for these two traces, we also observe that as we provide more learning data to the model predicts better i.e. prediction accuracy increases with increasing days of the week. For MACACO-Brazil traces, the prediction model is able to achieve accuracy between 75-80%. This happens because, in MACACO-Brazil trace, we were not able to get contact patterns for each day of the week due to missing contacts among people or due to the fact that people were not collecting data. From above results I derive that to provide a satisfactory prediction of time-varying pair-wise contact patterns, we need at least one week of learning data.

Finally, I also present the important top 20 features for both MIT and MACACO traces in Figure 4.6. From the feature rank table, we observe that a total number of contacts among the pair (in Figure, 1 represents the first person and 2 presents the second person of the pair) is the most important feature followed by the day of the week. Further, I also observe the contact features of each person in the pair are also crucial in predicting the time-varying contact probabilities. The standard deviation of inter-contact time between the pair and also with others in the network are also important features to enable better prediction accuracy of contact probabilities.

Finally, to cope up with the scalability issues while deploying this machine learning based approach, we can first prune those pairs who exhibits very fewer contacts among them. Further, we can also adopt incremental feature development by pre-processing the data in advance to make our model much faster and enable it to develop on-the-fly features. Finally, the model can be further fastened to handle the scalability concerns by only developing the features that came out as important features among different samples of datasets.

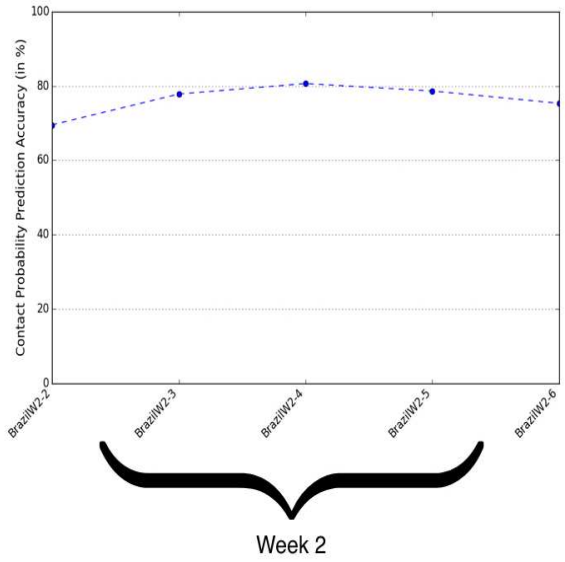
## 4.5 Conclusions

In this Chapter, I presented the *Contact Probability Prediction Module* of INDIGO. More specifically, I presented how the *Contact Probability Prediction Module* predicts the heterogeneous pair-wise contact probability for both static and time-varying contact patterns. For static contact patterns, I presented how can we




(a) MIT

(b) MACACO-France




(c) MACACO-Brazil

Figure 4.5. Mean Absolute Percentage Accuracy for predicted time-varying contact probabilities for different days of the different weeks of MIT, MACACO-France and MACAO-Brazil taces.

| Feature Name | Rank |
|---|---|
| TOTAL_CONTACTS_12 | 1 |
| DAY_NUM | 2 |
| MEAN_CONTACT_ALL_1 | 3 |
| MEAN_CONTACT_ALL_2 | 4 |
| MEDIAN_CONTACT_ALL_2 | 5 |
| TOTAL_CONTACT_ALL_1 | 6 |
| STD_INTER_CONTACT_TIME_ALL_2 | 7 |
| STD_INTER_CONTACT_TIME_ALL_1 | 8 |
| STD_INTER_CONTACT_TIME_12 | 9 |
| Q25_CONTACT_ALL_1 | 10 |
| TOTAL_CONTACTED_PEOPLE_1 | 11 |
| TOTAL_CONTACTS_ALL_2 | 12 |
| COFF_IC_TIME_ALPHA_12 | 13 |
| COFF_IC_TIME_XMIN_12 | 14 |
| Q75_INTER_CONTACT_TIME_ALL_2 | 15 |
| Q75_INTER_CONTACT_TIME_ALL_1 | 16 |
| Q25_CONTACT_ALL_2 | 17 |
| COFF_IC_TIME_ALL_ALPHA_2 | 18 |
| MEDIAN_CONTACT_ALL_1 | 19 |
| TOTAL_CONTACTED_PEOPLE_2 | 20 |

Figure 4.6. Rank of top 20 obtained from the feature selection using Recursive Feature Elimination.

utilize the inter-contact time distribution and Maximal Likelihood Estimation method like Conference and events in an Urban area. Further, I present the methodology to utilize Gradient Boosting Machine Learning Approach to learning the contact patterns of people and predict their time-varying day-wise contact probabilities automatically for contact traces with a longer duration like University environment. To enable time-varying contact probabilities prediction, I capture contact patterns of people in the form of features mainly related to contacts, the inter-contact time between a pair and all other people. I present the results for MIT, MACACO France, and MACACO Brazil traces and showed that my model is able to achieve a good accuracy with the incremental learning approach. I observe that as soon as we provide more learning data to the model, the prediction accuracy increases. I also summarize the key observation from this Chapter as follows:

- A Gradient Boosting Machine based increment learning approach can enable the prediction of future time-varying contact probabilities by learning

the past contact patterns of people.

- To get reasonable accuracy in contact probability prediction, the model requires longer contact traces with at least one week of data.
- The total number of contacts among the pair and day of the week contribute greatly for the time-varying contact probability prediction of a pair.
- The proposed time-varying contact pattern prediction provides more realistic contact patterns thus can empower INDIGO to provide much tighter and realistic upper bound of data dissemination time.

The work done in this Chapter has resulted in a publication at ACM MobiOpp 2012, ICT4S 2013 and ACM MobiCom 2015, IEEE Med-Hoc-Net 2015 conferences. In next Chapter, I will present the modeling part of INDIGO for Type II and Type III cases where social proximity dominates. I will present the modeling of data dissemination process under the broadcast data dissemination strategy and will also discuss the utilization of the contact probabilities predicted for both static and time-varying contact patterns from this Chapter. In next Chapter, I will also present the impact of time-varying contact pattern while predicting the upper bound of data dissemination time.

# Chapter 5

# Prediction of Upper Bound of Data Dissemination Time Under Broadcast Strategy

## 5.1 Introduction

In the previous Chapter, I described the prediction of heterogeneous pair-wise contact probabilities considering both static and time-varying nature of human mobility. In this Chapter, I will focus on the upper bound of data dissemination time for contact traces under broadcast data dissemination strategy for both static and time-varying pair-wise heterogeneous contact probabilities for both Type II. The broadcast data dissemination strategy under static contact probabilities is addressed in literature Peres et al. [2011] Picu et al. [2012] Mosk-Aoyama and Shah [2008] for a single data source. However, the existing works do not consider the time-varying nature of pair-wise contact probabilities and also do not take into account the impact of multiple simultaneous contacts and multiple data sources. In this Chapter, I will present a Markov chain based model called *DDT-Markov* of INDIGO framework that can realistically predict the upper bound data dissemination time of multi-contact and multi-source data dissemination under broadcast data dissemination strategy by considering the heterogeneous and time-varying mobility of people. The model takes the pair-wise contact probabilities among people as an input. I also present how the presented model *DDT-Markov* achieves much tighter and realistic upper bound than existing approaches by utilizing the exponential cut-off property of inter-contact time distribution that impacts the data gathering process Karagiannis et al. [2007] Garg et al. [2013].

The chapter is structured as follows. In Section 5.2, I will present the de-

tailed description of the components required to predict the upper bound of data dissemination time under the broadcast strategy. Further, in Section 5.3, I will present the modeling of multi-contact and multi-source data dissemination time under heterogeneous mobility and broadcast data dissemination strategy. In this Section, I will also present the first basic data dissemination algorithms to predict upper bound of data dissemination time using Markov model for both static and time-varying contact patterns. Further, Section 5.4 will present different approaches considered in this thesis for the tighter prediction of upper bounds of data dissemination time. After the explanation of the modeling and tighter bound approaches, I will present the methodology to create ground truth data dissemination time and the results obtained from different contact traces in Section 5.5. Section 5.6 will present the results obtained for the upper bound of data dissemination time for contact traces from the diverse environment and duration and also compare our solution with the key existing methods. Finally, I will conclude the Chapter with Section 5.7.

## 5.2 Overview of INDIGO framework components required under broadcast strategy

Figure 5.1 outlines the enlarged view of different components of INDIGO framework required to predict upper bound of data dissemination time under broadcast data dissemination strategy for Type II case. To predict the upper bound of data dissemination time under broadcast strategy, INDIGO considers *Contact Probability Prediction Module* and *DDT-Markov Component* of `Broadcast Sub-Module`. As explained in previous Chapter 4, the *Contact Probability Prediction Module* predicts the static and time-varying pair-wise heterogeneous contact probabilities among people using and/or learning their contact patterns from the real world or synthetic traces. After the contact probabilities are predicted, the INDIGO input them to the *DDT-Markov Component*. With the help of these modules and components, INDIGO framework predicts the upper bound of data dissemination time utilizing the Markov chain based model. Table 5.1 presents different notations used in this Chapter.

### 5.2.1 Contact Probability Prediction Module

Utilizing the real/synthetic contact traces and *Contact Probability Type* parameter (0 for static contact probabilities and 1 for time-varying contact probabilities), this module predicts the contact probabilities among each pair of people $i$ and

Table 5.1. Notations used in INDIGO under broadcast data dissemination strategy for Type II case.

| Notations | Description |
|---|---|
| $U, D$ | set of mobile users, set of mobile users selected as data sources |
| $N, M$ | number of mobile users, number of data sources |
| $u_j, d_i$ | mobile user $u_j \in U$, data source $d_i \in D$ |
| $G$ | graph of mobile users $U$ and data sources $D$ |
| $V$ | $V = U \cup D$, set of all mobile users and data sources |
| $p_{ij}^c$ | static contact probability between any pair $i, j \in V$, $p_{ij}^c \in [0,1]$ |
| $p_{ij}^{c_{day_k}}$ | contact probability between any pair $i, j \in V$, $p_{ij}^{c_{day_k}} \in [0,1]$ for $kth$ day |
| $P_S^C$ | contact probability matrix of all $p_{ij}^c$ for static pair-wise contact probabilities |
| $P_{TV}^C$ | set of day-wise contact probability matrices of all $p_{ij}^c$ for time-varying pair-wise contact probabilities |
| $P_{day_k}^C$ | contact probability matrix of all $p_{ij}^{c_{day_k}}$ for $kth$ day |
| $\alpha$ | Cut-off point used by Markov model, $\alpha \in [0,1]$ |
| $\mathbb{D}$ | overall data collection requirement of all mobile users $\mathbb{D} \in [0,1]$ |
| $F$ | $F = M * N * \mathbb{D}$, final number of messages to collect |
| $msg_i$ | distinct message associated to data source $d_i \in D$ |
| $MList_{u_j}(t)$ | list of all data messages received upto time $t$ by mobile user $u_j$ |
| $D^{ALL}(t)$ | list of all data messages collected upto time $t$ by all $U$ mobile users |
| $DF(t)$ | fraction of data collected till time $t$, $DF(t) \in [0, F]$ |
| $S(x)$ | Markov model state with $x$ messages collected by all mobile users |
| $S(F)$ | Markov model's target state with $F$ messages |
| $P_{S(x)S(x+h)}$ | transition probability to reach $S(x+h)$ from $S(h)$ under static contact probabilities |
| $P_{S(x)S(x+k)}^k$ | transition probability to reach $S(x+h)$ from $S(h)$ under time-varying contact probabilities |
| $\Delta$ | time step size of Markov model |
| $T_x$ | maximum time spent in Markov state $S(x)$ |
| $T_{dss^B}^{upper}$ | predicted upper bound of data dissemination under broadcast data dissemination strategy |
| $T_{FGP^B}$ | maximum time spent Fast Growing Phase |
| $T_{LTP^B}$ | maximum time spent Long Tail Phase |
| $T_{dss^B}^{meas}$ | measured data dissemination time from real traces under broadcast data dissemination strategy |
| $t^-, t^+$ | time before and after any time $t$ |

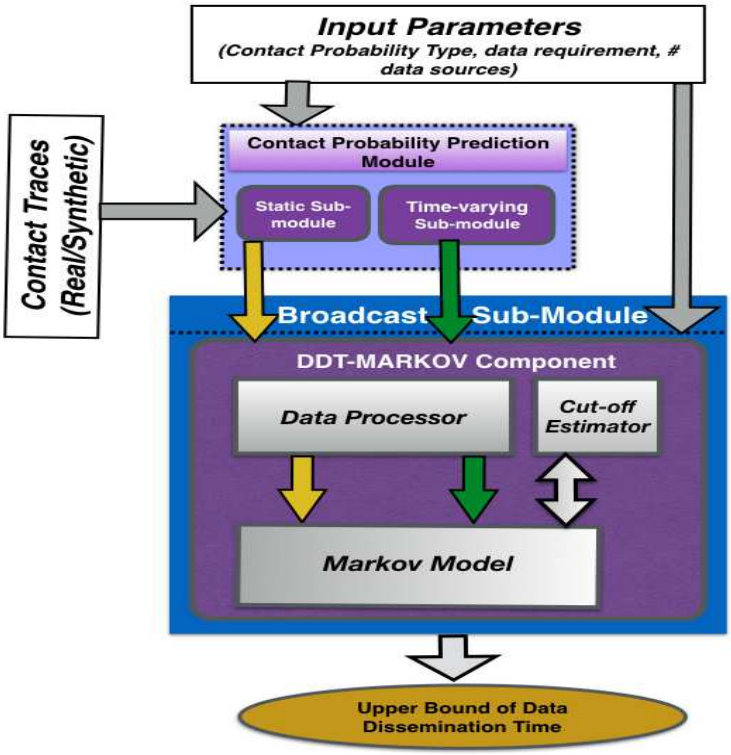



Figure 5.1. An enlarged view of the INDIGO components required to predict the upper bound of data dissemination time under broadcast data dissemination strategy for both static and time-varying pair-wise heterogeneous contact patterns.

$j$ . For static contact patterns (i.e. *Contact Probability Type* = 0) it constructs a contact probability matrix $P_S^C$ with static pair-wise heterogeneous contact probabilities $p_{ij}^c$ between each pair $i$ and $j$ using Equation 4.9 of Chapter 4. Further for time-varying contact patterns (i.e. *Contact Probability Type* = 1), it constructs a contact probability matrix set $P_{TV}^C$ that contains pair-wise heterogeneous contact probabilities for each pair in different days $p_{ij}^{c_{day_k}}$ (using Equation 4.9 of Chapter 4).

### 5.2.2 Broadcast Sub-Module

This sub-module come under the *Data Dissemination Prediction Module* of INDIGO framework (from Figure 3.2 of Chapter 3) and is responsible for predicting the tighter upper bound of data dissemination time for broadcast dissemination strategy. The core is the *DDT-Markov Component* that further consists of

Data Processor, Markov Model and Cut-off Estimator sub-components for the prediction. The Data Processor pre-processes different inputs required by the Markov Model and employs a Cut-off point based approach to predict tighter upper bound of data dissemination time for both static and time-varying contact patterns. The Markov Model communicates with Cut-off Estimator to estimate the Cut-off point $\alpha$ that plays a significant role to provide the tighter prediction of data dissemination time $T_{dss^B}^{upper}$ (more details of $\alpha$ is provided in next Sections). The Cut-off point based approach incorporates realistic aspects of human mobility and communication (i.e. heterogeneous contact patterns and multiple simultaneous contacts among people) and multiple data sources.

### Data Processor

This sub-component processes all inputs coming from *Contact Probability Prediction Module* for both static and time-varying pair-wise contact probabilities. It also uses *Input Parameters* given to the INDIGO framework. The main parameter for broadcast data dissemination strategy is the # of data sources that enables multi-source data dissemination. Out of all users, the Data Processor randomly assigns $M$ users as data sources and marks the rest as $N$ mobile users. Based on the type of contact probabilities (i.e. static or time-varying), it gives either $P_S^C$ contact probability matrix or $P_{TV}^C$ contact probability matrix set as an input to the Markov Model. These contact probability matrix (or matrix set) drives the mobility and heterogeneous contact patterns for the *DDT-Markov Component*.

### Markov Model

It is a Markov-chain based model that utilizes a Cut-off point based approach to mimic real-world data gathering process by considering the heterogeneous mobility patterns of people, multiple simultaneous contacts among people and data originating from multiple data sources. The Markov Model communicates with Cut-off Estimator to dynamically estimate the Cut-off point for the tighter prediction of the upper bound of data dissemination time. Each state $S(x)$ of the Markov Model represents the total number of messages $(x)$ collected by mobile users from different data sources. In this way, the Markov Model keeps on transiting from lower states to higher ones based on the number of messages collected in the network. The transition from one state to another state is driven by the mobility and contact patterns of mobile users and data sources. Once all mobile nodes collect all messages in the network then the Markov Model stops transiting and remain in the same state also called as an absorbing state. Due

to this absorbing state, this Markov Model is also called an Absorbing Markov Chain-based model. The detailed description of data dissemination modeling using Markov Model is presented in Section 5.3.

Cut-off Estimator

This component is also an integral part of INDIGO framework that contributes towards the tighter prediction of the upper bound of data dissemination time by dynamically providing the Cut-off point $\alpha$ value to Markov Model (more detail in later sections). To calculate $\alpha$, Markov Model communicates with Cut-off Estimator by providing the fraction of data collected $DF(t) \in [0,1]$ at time $t$ in the network at any time step $t$. Based on this information Cut-off Estimator measures the change in data fraction between two consecutive steps of Markov chain and repeats this process until it reaches to a data fraction $\alpha$ beyond which change in $t$ results in data fraction changes smaller than $\epsilon$. For any data fraction $DF(t) \in [0,1]$ at time $t$ and time step size $\Delta$, $\alpha$ can be calculated as:

$$\alpha : \frac{DF(t+\Delta) - DF(t)}{\Delta} << \epsilon, \epsilon = 10^{-4} \tag{5.1}$$

## 5.3 Modeling of multi-contact multi-source data dissemination using Markov chains under broadcast strategy

In this section, I present the modeling of multi-contact multi-source data dissemination process for both static and time-varying contact probabilities using Markov chains from Markov Model. Let us consider the entire network as a graph $G = (V, E)$ with $V = U \cup D$ where $U = \{u_1, u_2, ....., u_N\}$ represents $N$ mobile users and $D = \{d_1, d_2, ....., d_M\}$ represents $M$ data sources (mobile or static). An edge $(y, z) \in E$ between any two nodes $y \in V$ and $z \in V$ exists if (and only if) they can communicate with each other [1]. I assume that every data source $d_i \in D$ has a distinct data message $msg_i$. Further, for broadcast data dissemination strategy, I assume that each mobile user $u_j \in U$ is interested in gathering messages from all data sources. Therefore, the maximum number of messages that can be stored in the network are $F = M * N$ (each one of the $N$ users could have $M$ data messages). Every mobile user $u_j$ maintains a list $MList_{u_j}(t) = \{msg_i, i \in [1, M]\}$ of

[1] We assume that the communication between $y$ and $z$ is bi-directional.

all messages it receives up to time $t$. All mobile users of network $G$ collect data in two possible ways (or processes):

- *User–User Contact (UUC):* When any two mobile users come in contact with each other and exchange their respective message lists. This process is similar to epidemic routing.
- *User–Data Source Contact (UDSC):* When a mobile user encounters a data source and collects its data message.

**Algorithm 1** Data gathering algorithm via *UUC* process

1: **if** Any two users $u_j$ and $u_k$ come in contact at time $t$ with $MList_{u_j}(t^-)$ and $MList_{u_k}(t^-)$ **then**
2: Users $u_j$ and $u_k$ exchange all of their data messages
3: $MList_{u_j}(t^+) = MList_{u_j}(t^-) \cup MList_{u_k}(t^-)$
4: $MList_{u_k}(t^+) = MList_{u_j}(t^-) \cup MList_{u_k}(t^-)$
5: **end if**

**Algorithm 2** Data gathering algorithm via *UDSC* process

1: **if** Any user $u_j$ with $MList_{u_j}(t^-)$ and any data source $d_i$ with message $msg_i$ come in contact at time $t$ **then**
2: Users $u_j$ collects data message $msg_i$ from $d_i$
3: $MList_{u_j}(t^+) = MList_{u_j}(t^-) \cup msg_i$
4: **end if**

It is important to note that any two data sources *do not* exchange any data messages among them. I define multi-contact multi-source data dissemination process as follows:

*Definition 1 (Multi-contact Multi-source Data Dissemination under Broadcast): Given a network G, any user $u_j \in U$ gathers data message $msg_i$ from data source $d_i \in D$ or further disseminates its messages to other users via UUC process. Any user $u_k \in U$ can also directly receive data message $msg_i$ using UDSC process. The data dissemination process continues until all mobile users in the network collect data from all data sources. It is driven by the mobility of users (and data sources) and multiple simultaneous contacts among them.*

Algorithm 1 and Algorithm 2 present how every user $u_j \in U$ gathers data at time $t$ using *UUC* and *UDSC* processes. $t^-$ and $t^+$ represent the time before and after $t$ respectively.

### 5.3.1 Preliminaries and assumptions

In order to model the mobility of people, the Markov Model utilizes the processed contact probabilities matrix obtained from Data Processor for all mobile nodes and mobile nodes and data sources. Please note that the predicted contact probabilities were provided by the *Contact Probability Prediction Module* for either static or time-varying contact pattern settings. In case of static contact patterns, the contact probability matrix is $P_S^C$ while in case of time-varying contact probabilities, it is given as a set of contact probability matrix $P_{TV}^C$ consisting of contact probability matrix for $k$ days i.e. $P_{TV}^C = \{P_{day_1}^C, P_{day_2}^C ..... P_{day_k}^C\}$ where $k \in [1, 7]$ respectively (details were provided in Chapter 4).

To model multiple simultaneous contacts, the Markov Model uses a synchronous time model, where time is slotted commonly across all nodes in the network. In any time slot, each node in the network $G$ may contact one of its neighbors according to a random choice that is independent of the choices made by other nodes. Thus, in synchronous time model, multiple contacts may occur simultaneously Mosk-Aoyama and Shah [2008]. Before giving a detailed description of the modeling of data dissemination time using the Markov Model, I will first present some of the assumptions taken during the modeling process and the definition of data dissemination time under broadcast data dissemination strategy.

**Assumptions:**

- The contact probability matrix $P_S^C$ or any $P_{day_k}^C$ is doubly stochastic matrix i.e $p_{ij}^c = p_{ji}^c \; i, j \in V$.
- Number of data sources are very less compared to a number of users[2] i.e $M << N$.
- Data dissemination time is finite.
- Similar to existing works, I also assume that the duration of contact is small but sufficient to transfer all data Gao and Cao [2011] Li et al. [2013].

*Definition 2 (Data Dissemination Time Upper Bound Under Broadcast): I define data dissemination time as the time at which all users in the network receive all (or maximum feasible[3]) messages from all data sources.*

[2]This assumption is admissible because in a real world scenario, generally, the number of data sources are less as compared to receivers.

Consider a matrix $D^{ALL}(t)$ of size $N \times M$ that represents a list of all data messages collected up to time $t$ by $N$ mobile users.

$$D^{ALL}(t) = \begin{bmatrix} msg_{u_1 d_1} & msg_{u_1 d_2} & \dots & msg_{u_1 d_M} \\ msg_{u_2 d_1} & msg_{u_2 d_2} & \dots & msg_{u_2 d_M} \\ . & . & . & . \\ . & . & . & . \\ msg_{u_N d_1} & msg_{u_N d_2} & \dots & msg_{u_N d_M} \end{bmatrix}_{N \times M}$$

$$msg_{u_j d_i} = \begin{cases} 1 & if \quad msg_i \in MList_{u_j}(t) \\ 0 & \text{otherwise} \end{cases}$$

$$\forall\ j \in [1, N],\ \forall\ i \in [1, M]$$

The upper bound of data dissemination time under broadcast strategy $T_{dss^B}^{upper}$ is the maximum time slot at which data requirement will be fulfilled and $F$ elements of matrix $D^{ALL}(t)$ become 1. It is important to note that since I am using Markov chain based model, the $T_{dss^B}^{upper}$ will be expressed in event time (number of contact events or ticks) as opposed to standard wall clock time. I convert this event time to wall clock time by multiplying it with the beacon interval.[4]

### 5.3.2 Prediction of $T_{dss^B}^{upper}$ using Markov model

Each state $S(x), x \in [0, F]$ of the Markov Model represents the total number of messages $(x)$ collected by mobile users either from different data sources or through other mobile nodes and can be viewed as data matrix $D^{ALL}(t)$ at time $t$, where the number of non-zero elements represent the number of messages present in a network (starting from 0 to $F$ where $F = MN$ represents the target state to reach). Figure 5.2 represents one realization of the Markov chain[5]. The state transition models the increase in the number of messages in the network and is driven by the probability of transition from one state to another. The probability of transition $P_{S(x),S(x+h)}$ from any state $S(x)$ to $S(x+h)$ can be calculated using contact probability matrix $P_S^C$ for static pair-wise contact probabilities while contact probability matrix set $P_{TV}^C$ for time-varying contact probabilities. The transition probability for both cases is defined in Equation 5.2 and 5.3 respectively.

[4] Based on scan interval in contact trace.

[5] There could be several realizations of one state for example when there is 1 message in the network then it can start with any user and there could be $\binom{MN}{1}$ possible realizations.

$$P_{S(x)S(x+h)} = \sum_{i \in S(x), j \in S(x+h)} p_{ij}^{c} \tag{5.2}$$

$$P_{S(x)S(x+h)}^{k} = \sum_{i \in S(x), j \in S(x+h)} p_{ij}^{c_{day_k}} \tag{5.3}$$

$$\text{Where } \forall\ p_{ij}^{c} \in P^{C},\ \forall\ p_{ij}^{c_{day_k}} \in P_{day_k}^{C},\ \forall\ P_{day_k}^{C} \in P_{TV}^{C}$$

Where $P_{S(x)S(x+h)}$ presents the transition probability to reach state $S(x)$ to $S(x+h)$ during static contact patterns case while $P_{S(x)S(x+h)}^{k}$ corresponds to the transition probability to reach state $S(x)$ to $S(x+h)$ for any $kth$ day i.e. $day_k, k \in [1,7]$ in time-varying contact patterns case. Please note that $P_{S(x)S(x+k)}^{k}$ will vary according to the predicted contact probabilities at different days obtained through the *Contact Probability Prediction Module*. Once we reach the final state $S(F)$, the transition probability to remain in the same state will be 1 (also called absorbing state). Figure 5.3 represents the sample probability transition matrix $P$ for $N$ mobile nodes, where each state represents a total number of messages in the network and $S(F)$ presents the final state of Markov chain.

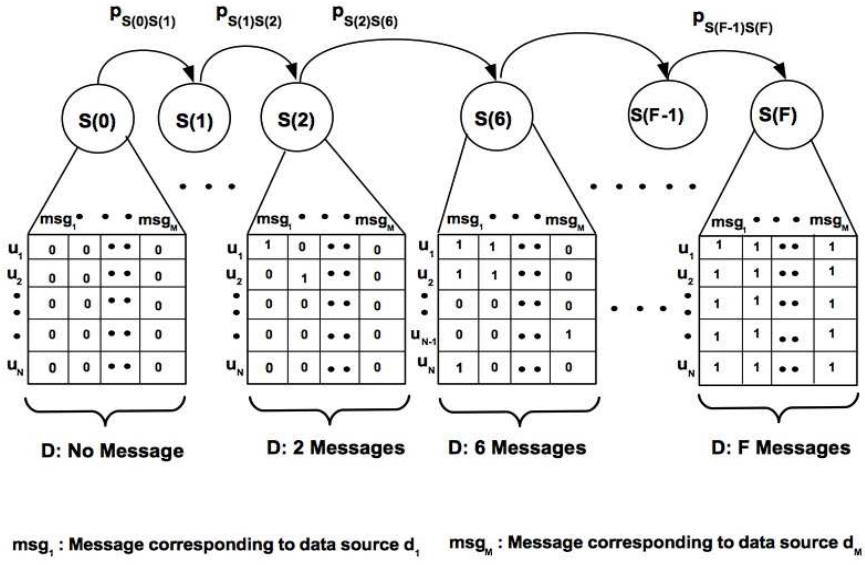



Figure 5.2. One realization of `Markov Model` starting from $S(0)$ to $S(F)$. This realization also shows the jump from $S(2)$ to $S(6)$ due to simultaneous contact between two mobile users and mobile user and data sources.

Further, the upper bound of data dissemination time $T_{dss^B}^{upper}$ can be approximated as the total time spent in each state before reaching the final state $S(F)$ in

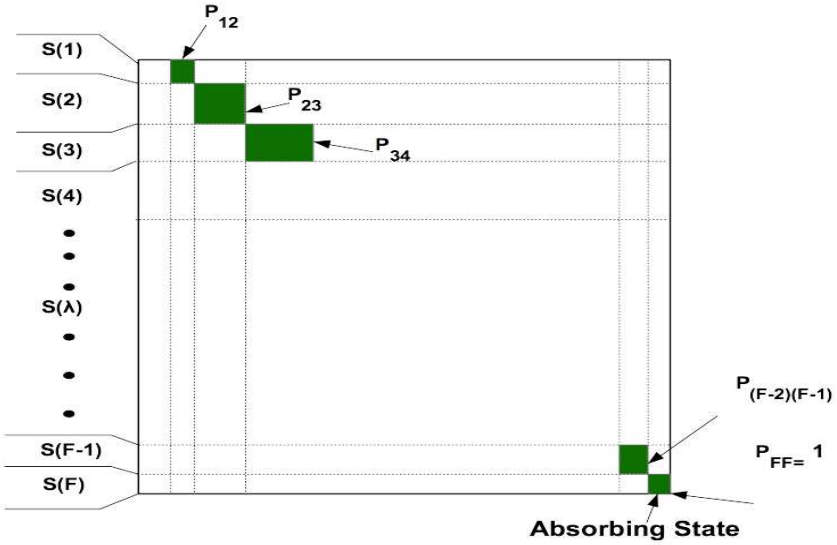



Figure 5.3. The sample transition probability matrix $P$ for our `Markov Model` where $M$ and $N$ are the total number of data sources and mobile nodes respectively.

the network. Let $T_x$ be the time spent in each state $S(x)$ before the transition to any other state. Finally, the $T_{dss^B}^{upper}$ can be predicted as follows:

$$T_{dss^B}^{upper} = \sum_{x=0}^{F-1} T_x \tag{5.4}$$

Due to multiple simultaneous contacts and multiple data messages, this Markov Model is not always a 1-step transition chain. For example, let us assume that we are in state $S(2)$ of Figure 5.2 where two users $u_1$ and $u_2$ have one unique data message in their respective message list ($MList_{u_1} = \{msg_1\}$ and $MList_{u_2} = \{msg_2\}$). When they come in contact, message lists $MList_{u_1}$ and $MList_{u_2}$ will be exchanged. Similarly, at the same time users, $u_{N-1}$ and $u_N$ come in contact with two data sources and update their message lists as $MList_{u_{N-1}} = \{msg_M\}$ and $MList_{u_N} = \{msg_1\}$ respectively. In this case, we directly jump to state $S(6)$ as the total number of messages in the network becomes 6. Therefore, in this case, the time spent in $S(3)$, $S(4)$ and $S(5)$ states is zero. To avoid an overestimated value of $T_{dss^B}^{upper}$, we need not count the time spent in each of these three states. Thus, multiple simultaneous contacts and multiple data messages from multiple data sources[6] can lead to skipping some states. Therefore, my model finds out

[6]this is what happens in real scenarios

and eliminates those states that are never reached (or finds multi-step transition of Markov chains).

I will now present the working of basic data dissemination algorithm to predict $T_{dss^B}^{upper}$ for both static and time-varying contact patterns. The basic algorithm is similar in both cases however the only difference comes the way contact probabilities are utilized inside the Algorithm.

$T_{dss^B}^{upper}$ prediction–Static pair-wise contact patterns case:

For the static case, I considered that each pair of people will have the heterogeneous fixed contact probability across different time periods. In this case, the Markov Model only takes a single contact probability matrix $P_S^C$ as an input to predict the upper bound of data dissemination time, $T_{dss^B}^{upper}$. Figure 5.4 presents the process to predict $T_{dss^B}^{upper}$ under static contact patterns case. Please note that static contact probability is the widely used notion of considering mobility patterns of people in literature Peres et al. [2011] Picu et al. [2012] Mosk-Aoyama and Shah [2008].

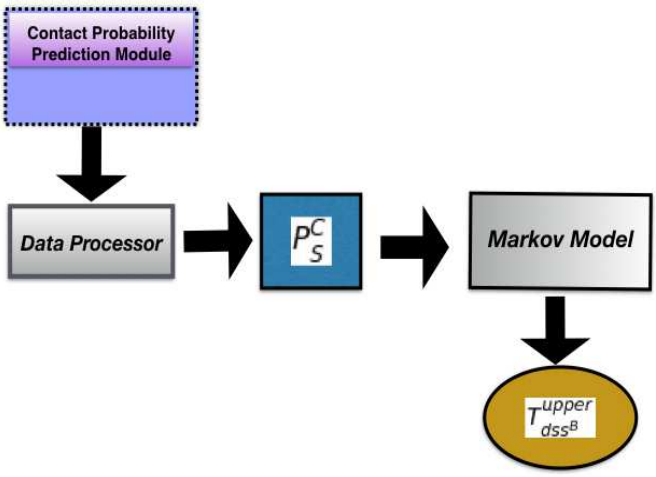



Figure 5.4. The process to predict the upper bound of data dissemination time under static pair-wise contact probabilities.

Utilizing Equation 5.4, I approximate Markov chain and present the basic algorithm to predict the upper bound of data dissemination time $T_{dss^B}^{upper}$ in Algorithm 3. In Algorithm 3, I present the method to find total number of steps required to compute $T_{dss^B}^{upper}$ in a network $G$ consisting of $N$ mobile users and $M$ data sources. In the static case, we see that the contact probability between any pair of people or data source remains static over time. Further, Algorithm 4

**Algorithm 3** Basic Data Dissemination Time algorithm static case

**Require:** $P_S^C$, $M > 0$, $N > 0$, $D^{ALL}(t) = 0_{N \times M}$, $MaxRuns$,$F$
**Ensure:** $T_{dss^B}^{upper}$
  $startState \leftarrow 0$
  $stopState \leftarrow F$
  $T_{dss^B}^{upper} \leftarrow 0$
  **while** $startState < stopState$ **do**
    $T_{runs} = \emptyset$
    $nextState \leftarrow startState + 1$
    **for** $run = 1$ **to** $MaxRuns$ **do**
      $p \leftarrow startState$
      randomly set $p$ elements of matrix $D^{ALL}(t)$ to 1
      $maxTime \leftarrow getMaxTimeStatic(startState, nextState, D^{ALL}(t), P_S^C)$
      $T_{runs} \leftarrow T_{runs} \cup maxTime$
      **if** $maxTime = 0$ **then**
        $nextState \leftarrow nextState + 1$
      **else**
        $startState \leftarrow nextState$
        $nextState \leftarrow nextState + 1$
      **end if**
    **end for**
    $T_{dss^B}^{upper} \leftarrow T_{dss^B} + max(T_{runs})$
  **end while**

presents maximum time spent in one state i.e time taken to reach $nextState$ from $startState$ required by my basic algorithm. Initially, my model set $S(0)$ as $startState$ and utilizes $getMaxTimeStatic$ method utilized in Algorithm 4 to compute maximum time required to reach $nextstate$ $S(1)$ for several trials i.e. $MaxRuns$ (for the better approximation of time spent in each state). If it cannot reach $S(1)$ due to the non-existence of *UUC* and *UDSC* processes for $MaxRuns$ then, it considers this state as non-reaching and set $maxTime$ to reach $S(1)$ from $S(0)$ as 0. Further, it sets $S(2)$ as $nextstate$ and computes $maxTime$ to reach $S(2)$ from $S(0)$. Otherwise, if it could reach $nextstate$ $S(1)$ (i.e $maxTime \neq 0$) then, it set $startState$ as $S(1)$ and next state as $S(2)$). Likewise, it repeats the same process until it reaches the final state $S(F)$. *UUC* and *UDSC* processes are modeled using the static pair-wise contact probabilities of matrix $P_S^C$. Utilizing $P_S^C$, $startState$, and $nextState$, I find $maxTime$ spent in each state using the $getMaxTimeStatic$ method and finally predict the upper bound of data dissemination time $T_{dss^B}^{upper}$.

### $T_{dss^B}^{upper}$ prediction–Time-varying pair-wise contact patterns case:

In this section, I present how to predict the upper bound of data dissemination time when we adopt a realistic aspect of mobility i.e. time-varying contact patterns. For time-varying contact patterns, I utilized contact probability matrix set $P_{TV}^C = P_{day_1}^C, P_{day_2}^C ..... P_{day_k}^C$ where $k \in [1, 7]$ given as input through the Data Processor of `Broadcast Sub-Module`. As we have seen in Chapter 4 that prediction of time-varying contact pattern is significant in those contact traces that exhibits regular mobility patterns of people over time and collected over a long duration of time. We also saw in the previous Chapter that we require at least one week of data to train the model and to obtain good accuracy for contact probabilities. Based on these observations in Chapter 4, I utilize daily-wise contact probabilities to predict $T_{dss^B}^{upper}$ for time-varying contact patterns.

To the best of my knowledge, automatic learning of time-varying contact pattern (through the *Contact Probability Prediction Module*) and their utilization in data dissemination process is not taken into account in literature. To address this issue, I further improve my basic algorithm (Algorithm 3) to model time-varying contact patterns and predict the upper bound of data dissemination time, $T_{dss^B}^{upper}$. Figure 5.5 presents the process to predict $T_{dss^B}^{upper}$ under time-varying contact probabilities case. Once the day-wise contact probabilities are predicted through *Contact Probability Prediction Module* for all nodes in the network, they are given as an input to the Data Processor that further processes these contact probabilities for mobile nodes and data sources separately. Afterward, the Data Processor

**Algorithm 4** $getMaxTimeStatic(startState, endState, D^{ALL}(t), P_S^C)$

**Require:** $startState$, $endState$, $D^{ALL}(t)$, $P_S^C$, $MaxTrials$
**Ensure:** $maxTime$

**for** $trial = 1$ **to** $MaxTrials$ **do**
  $T = \emptyset$
  $t = 0$
  **repeat**
    $x \leftarrow \mathcal{U}(0,1)$
    $t \leftarrow t + 1$
    {Assuming *UUC* process}
    **if** $x < 0.5$ **then**
      $r \leftarrow \mathcal{U}(0,1)$
      **for all** pairs of users $(i,j) : p_{i,j}^c \geq r$ **do**
        update $D^{ALL}(t)$ based on $UUC$ algorithm
      **end for**
    **end if**
    {Assuming *UDSC* process}
    **if** $x > 0.5$ **then**
      $r \leftarrow \mathcal{U}(0,1)$
      **for all** user and data source pair $(i,k) : p_{i,k}^c \geq r$ **do**
        update $D^{ALL}(t)$ based on $UDSC$ algorithm
      **end for**
    **end if**
    $d \leftarrow count(D^{ALL}(t))$ {total messages stored in $D^{ALL}(t)$}
  **until** $d \leq endState$
  **if** $d = endState$ **then**
    $T \leftarrow T \cup t$
  **end if**
**end for**
$maxTime \leftarrow max(T)$
**return** $maxTime$

passes the set of probability matrix $P_{TV}^{C}$ to the Markov Model.

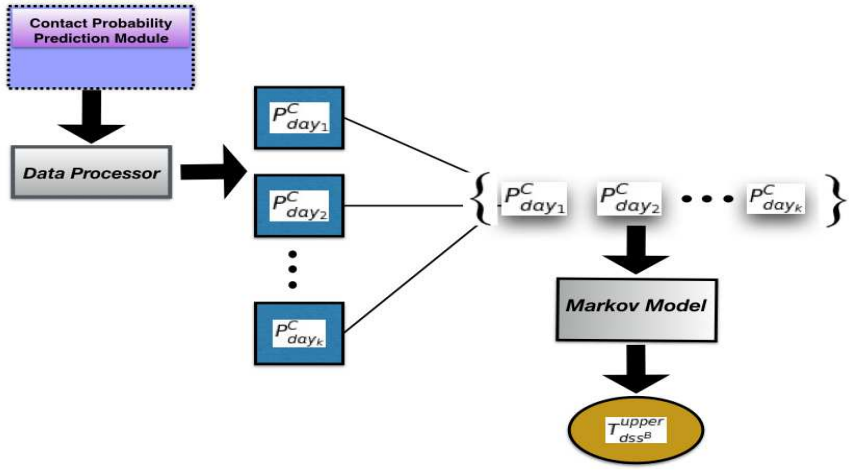



Figure 5.5. The process to predict upper bound of data dissemination time under time-varying pair-wise contact probabilities.

Algorithm 5 predicts $T_{dss^B}^{upper}$ under time-varying contact probabilities and Algorithm 6 computes the maximum time spent in each state while considering the time-varying contact probabilities i.e. max time taken to reach $nextState$ from $startState$. Similar to static case, Algorithm 5 also set $S(x)$ as $startState$ and utilizes $getMaxTimeTV$ method to compute maximum time required to reach $nextstate$ $S(x+1)$ for several trials i.e. $MaxRuns$. In the time-varying case, my model sends the set of contact probability matrix $P_{TV}^{C}$ and the $MaxStepsDay$. The $MaxStepsDay$ is the maximum steps that can be taken in a day by the Markov Model and can be calculated as:

$$MaxStepsDay = \lceil \frac{24 * 60 * 60}{\Delta} \rceil \tag{5.5}$$

Where $\Delta$ is the step size of Markov chain determined by the scan interval of the mobile devices. Using these parameters obtained from Algorithm 5 and *UUC* and *UDSC* processes, the Algorithm 6 finds the maximum time spent in the given state $S(x)$. To do this, the algorithm first finds the correct contact probability matrix $P^{C}_{day_{dayNum}}$ for the given $dayNum$ from $P_{TV}^{C}$. Afterward utilizing the specific day contact probability matrix the algorithm predicts $maxTime$ spend in a given state by using the contact probabilities between any pair of two mobile users or a pair of mobile user and data source. It is also important to note that if the total number of steps spent in a given state exceeds the total number of possible steps i.e. $MaxStepsDay \times \left|P_{TV}^{C}\right|$ then, our Markov Model restarts from

the first-day contact probability matrix. Similar to the static case, I also find the $maxTime$ by eliminating those steps that can never be reached. Likewise, the algorithm repeats the same process utilizing $P_{TV}^{C}$, $startState$, and $nextState$ and find $maxTime$ spent in each state using $getMaxTimeTV$ method until it reaches the final state $S(F)$. Finally, Algorithm 5 predicts $T_{dss^B}^{upper}$ by summing up the max time spent in each reachable state.

**Algorithm 5** Basic Data Dissemination Time algorithm time-varying case

**Require:** $P_{TV}^{C}$, $M > 0$, $N > 0$, $D^{ALL}(t) = 0_{N \times M}$, $MaxRuns$,$MaxStepsDay$,$F$
**Ensure:** $T_{dss^B}^{upper}$
  $startState \leftarrow 0$
  $stopState \leftarrow F$
  $T_{dss^B}^{upper} \leftarrow 0$
  **while** $startState < stopState$ **do**
    $T_{runs} = \emptyset$
    $nextState \leftarrow startState + 1$
    **for** $run = 1$ **to** $MaxRuns$ **do**
      $p \leftarrow startState$
      randomly set $p$ elements of matrix $D^{ALL}(t)$ to 1
      $maxTime \leftarrow getMaxTimeTV(startState, nextState, D^{ALL}(t), P_{TV}^{C}, MaxStepsDay)$

      $T_{runs} \leftarrow T_{runs} \cup maxTime$
      **if** $maxTime = 0$ **then**
        $nextState \leftarrow nextState + 1$
      **else**
        $startState \leftarrow nextState$
        $nextState \leftarrow nextState + 1$
      **end if**
    **end for**
    $T_{dss^B}^{upper} \leftarrow T_{dss^B} + max(T_{runs})$
  **end while**

The algorithms (Algorithm 3 and Algorithm 5) discussed above for both static and time-varying contact pattern cases presents the basic data dissemination algorithm to predict the upper bound of data dissemination time by excluding the time spent in unreachable states. However, the value of $T_{dss^B}^{upper}$ obtained by the summation of maximum time spent in each state ($maxTime$) is not necessarily equal to the maximum time taken to reach $S(F)$ from $S(0)$. Therefore, both algorithms will overestimate $T_{dss^B}^{upper}$.

Since this thesis focuses on providing tighter upper bounds of $T_{dss^B}^{upper}$, there-

**Algorithm 6** $getMaxTimeTV(startState, endState, D^{ALL}(t), P^{C}_{TV}, MaxStepsDay)$

**Require:** $startState$, $endState$, $D^{ALL}(t)$, $P^{C}_{TV}$, $MaxTrials$,$dayNum$,$MaxStepsDay$
**Ensure:** $maxTime$

**for** $trial = 1$ **to** $MaxTrials$ **do**
  $T = \emptyset$
  $t = 0$
  $tlocal = 0$
  $dayNum = 0$
  **repeat**
    $x \leftarrow \mathcal{U}(0,1)$
    $t \leftarrow t + 1$
    {Restart from the beginning contact probability if we reach $MaxStepsDay$}
    **if** $tlocal > MaxStepsDay \times \left|P^{C}_{TV}\right|$ **then**
      $tlocal \leftarrow 0$
    **end if**
    $tlocal \leftarrow tlocal + 1$
    {Get $P^{C}_{day_{dayNum}}$ for the given day using $tlocal$ and $MaxStepsDay$}
    $P^{C}_{day_{dayNum}} \leftarrow P^{C}_{TV}[dayNum]$
    {Assuming *UUC* process}
    **if** $x < 0.5$ **then**
      $r \leftarrow \mathcal{U}(0,1)$
      **for all** pairs of users $(i,j) : p_{i,j}^{c_{dayNum}} \geq r$ **do**
        update $D^{ALL}(t)$ based on $UUC$ algorithm
      **end for**
    **end if**
    {Assuming *UDSC* process}
    **if** $x > 0.5$ **then**
      $r \leftarrow \mathcal{U}(0,1)$
      **for all** user and data source pair $(i,k) : p_{i,k}^{c_{dayNum}} \geq r$ **do**
        update $D^{ALL}(t)$ based on $UDSC$ algorithm
      **end for**
    **end if**
    $d \leftarrow count(D^{ALL}(t))$ {total messages stored in $D^{ALL}(t)$}
  **until** $d \leq endState$
  **if** $d = endState$ **then**
    $T \leftarrow T \cup t$
  **end if**
**end for**
$maxTime \leftarrow max(T)$
**return** $maxTime$

fore, I further improve the basic data dissemination algorithms by employing the observations obtained from the real-world data gathering process. In next Section, I will present different approaches in detail that can be used to provide a tighter prediction of the upper bound of data dissemination time under the Broadcast strategy.

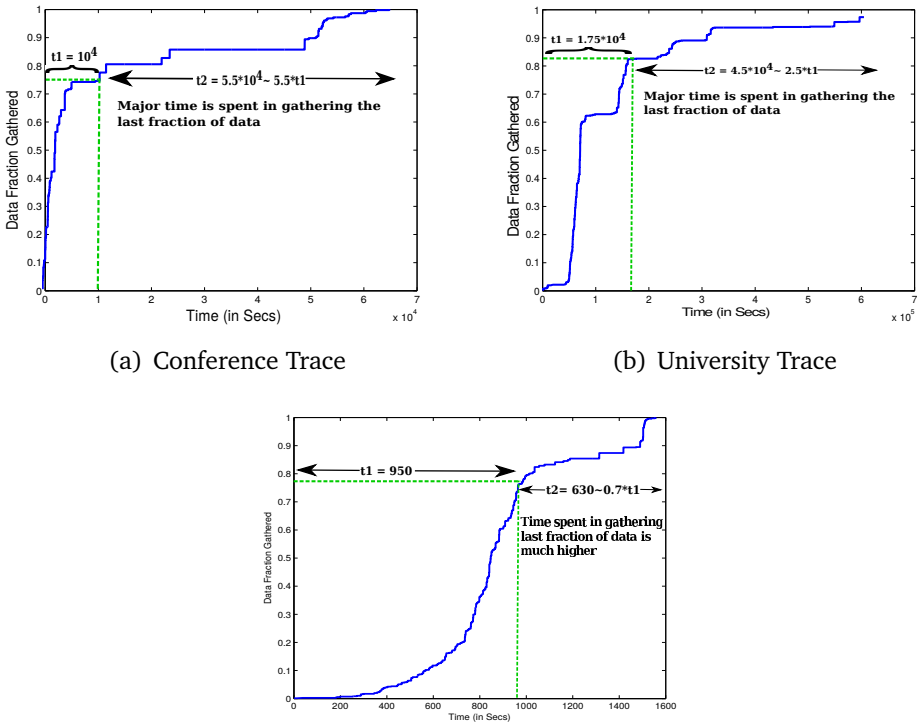
## 5.4 Tighter prediction of $T_{dss^B}^{upper}$

In order to achieve a much tighter upper bound for $T_{dss^B}^{upper}$, I analyze the fraction of data collected over time in different real-world traces. Figure 5.6 shows data gathering in three diverse environments Scott et al. [2006c] Eagle et al. [2009] Tournoux et al. [2009] under broadcast data dissemination strategy: *Conference*, *University*, and *Crowded Urban*. For all environment, we can observe that initially the rate of data gathering is very fast and after a certain data fraction, it exhibits a long tail cut-off. This happens because after a while the probability of getting unique messages from neighboring users reduces or some users have extremely low contact with rest of the users. Therefore, the time taken to gather remaining data increases significantly. In addition, the rate of data gathering depends on the inter-contact time among people that exhibit a long tail cut-off and also impacts the data gathering process Karagiannis et al. [2007]. I believe this property also impacts data gathering process. To predict much tighter upper bound for $T_{dss^B}^{upper}$, I utilize the above property of data gathering process and divide my analysis in 2-phases, based on the number of states $F$ (or $MN$).

1. **Fast Growing Phase:** In this phase, I find expected time required to reach a certain state $S(x)$ directly from $S(0)$ under broadcast data dissemination strategy. In this way, I try to mimic the initial fast rate of data gathering process. I call this expected time as $T_{FGP^B}$.

2. **Long Tail Phase:** For this phase, I find maximum time spent in each state starting from state $S(x)$ to final state $S(F)$. In this phase, I calculate maximum time spent in each state because time spent for collecting the last fraction of data has maximum weight on data dissemination time (as shown in Figure 5.6) and articulate long tail of data gathering process. I call this time as $T_{LTP^B}$.

$$T_{LTP^B} = \sum_{k=x}^{F} T_x \tag{5.6}$$

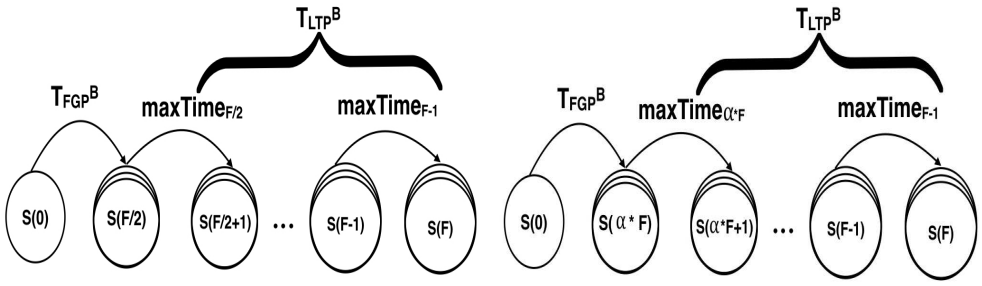



(a) Conference Trace

(b) University Trace

(c) Crowded Urban Trace

Figure 5.6. The Fraction of data gathered with respect to time for all mobile users from all data sources for real-world traces from three diverse environments under broadcast data dissemination strategy.




(a) Bisection approach where Fast Growing Phase starts from $S(0)$ to $S(F/2)$ and Long Tail Phase starts from $S(F/2)$ to $S(F)$.

(b) $\alpha$ Cut-off point approach where Fast Growing Phase starts from $S(0)$ to $S(\alpha \times F)$ and Long Tail Phase starts from $S(\alpha \times F)$ to $S(F)$.

Figure 5.7. Tighter prediction of the $T_{dss^B}^{upper}$ using Bisection and Cut-off point approach.

Finally, the upper bound of data dissemination time $T_{dss^B}^{upper}$ can be calculated as:

$$T_{dss^B}^{upper} = T_{FGP^B} + T_{LTP^B} \tag{5.7}$$

Now the important question arises is how do we determine the optimum state $S(x)$ that divides the **Fast Growing Phase** and **Long Tail Phase**. In order to set $S(x)$, I considered following two approaches to predict $T_{dss^B}^{upper}$ is shown in Figure 5.7(a) and Figure 5.7(b) respectively.

- Bisection approach
- Cut-off point approach

### Bisection approach

As the name suggests, this approach assumes that the data dissemination process is equally divided in Fast Growing and Long Tail Phase and is also used in literature to theoretically study single source data dissemination time Mosk-Aoyama and Shah [2008]. In this approach, the Fast Growing Phase goes from state $S(0)$ to $S(F/2)$ and Long Tail Phase from $S(F/2)$ to $S(F)$ where $S(F)$ presents the target state. In Fast Growing Phase rapid data collection results in higher number of messages in the network at each time slot. This phase utilizes Algorithm 3 and Algorithm 5 to predict $T_{FGP^B}$ as the time to reach $S(F/2)$ directly from $S(0)$ ($startState \leftarrow S(0)$ and $endState \leftarrow S(F/2)$) for static and time-varying contact patterns respectively. Further, $T_{LTP^B}$ (the time spent by network $G$ in Long

Tail Phase) that significantly impacts data dissemination time is estimated using Equation 6.4. For time spent in each of the $F/2$ states starting from $S(F/2)$ to $S(F)$ of Long Tail Phase, I calculate $maxTime$ using Algorithm 4 and Algorithm 6 ($startState \leftarrow S(F/2)$ and $stopState \leftarrow S(F)$) respectively for both cases.

### Cut-off point approach

The Bisection approach described in the previous section provides much tighter bound of $T^{upper}_{dss^B}$ as compared to basic data dissemination algorithm however it is not necessarily true that the Long Tail phase will always start after 50% of data collection. Therefore, to further articulate the realistic data gathering process (observed from Figure 5.6) and to further tighten the upper bound, I utilize the impact of long tail cut-off on data dissemination time by introducing the Cut-off point. As seen in Figure 5.6, long tail cut-off contributes higher weight on $T^{upper}_{dss^B}$ as compared to the time spend in gathering initial data messages (almost 5.5 times (INFOCOM), 2.5 times (MIT) and 0.7 times in Rollernet as compared to the time to collect initial data fraction). I believe that the consideration of this long tail cut-off in our model can provide a much tighter bound of $T^{upper}_{dss^B}$ because it closely articulates the real-world data gathering process. For this reason, the Markov Model utilizes a Cut-off point based approach and define a Cut-off Point $\alpha$ as follows (also described in Equation 5.1):

*Definition 3 (Cut-off Point $\alpha$): It is a Data Fraction point $DF(t) \in [0, 1]$ at time $t$ beyond which the change in data collection becomes smaller than a very small value $\epsilon$ and exhibits a long tail over time.*

To determine $\alpha$, the proposed Markov Model communicates with Cut-off Estimator of the INDIGO framework by sending the fraction of data collected in current and previous step (for more details refer Cut-off Estimator of Section 5.2) and learns it automatically. Once $\alpha$ is determined then, I predict $T^{upper}_{dss^B}$ using Equation 6.3. The Fast Growing Phase considers all states till cut-off point $\alpha$ i.e. from $S(0)$ to $S(\alpha \times F)$ and Long Tail phase takes it from $S(\alpha \times F)$ to $S(F)$. Algorithm 7 presents my modified algorithm to predict tighter upper bound for $T^{upper}_{dss^B}$ by computing $T_{FGP^B}$ and $T_{LTP^B}$ individually for both static and time-varying contact patterns using Algorithm 4 and Algorithm 6 respectively. This algorithm can also be used for Bisection approach by setting $\alpha$ as 0.5.

Once the Markov Model determines $\alpha$ through Cut-off Estimator, it segregates Fast Growing and Long Tail phase. In Fast Growing phase, the model sets

**Algorithm 7** Cut-off point based approach for prediction of $T_{dss^B}^{upper}$ for both static and time-varying contact patterns

**Require:** $P_S^C$, $P_{TV}^C$, $M > 0$,$N > 0$,$D^{ALL}(t) = 0_{N \times M}$, $NumRuns$,$F$,$MaxStepsDay$
**Ensure:** $T_{dss^B}^{upper}$
  $startState \leftarrow 0$
  $stopState \leftarrow F$
  $cutoffState \leftarrow 0$
  $T_{FGP^B} \leftarrow 0$, $T_{LTP^B} \leftarrow 0$, $T_{dss^B}^{upper} \leftarrow 0$, $\alpha \leftarrow 0$
  $T_{runs} = \emptyset$
  {estimate $\alpha$ from Cut-off Estimator sub-module}
  Set Cut-off point $\alpha$
  $cutoffState \leftarrow \alpha * F$ {calculate time for Fast Growing phase}
  **for** $i = 0$ **to** $NumRuns$ **do**
    **if** $STATIC\_CP\_CASE$ **then**
      $maxTime \leftarrow getMaxTimeStatic(startState, cutoffState, D^{ALL}(t), P^C)$ {For static contact patterns}
    **else**
      $maxTime \leftarrow getMaxTimeTV(startState, cutoffState, D^{ALL}(t), P_{TV}^C, MaxStepsDay)$ {For time-varying contact patterns}
    **end if**
    $T_{runs} \leftarrow T_{runs} \cup maxTime$
  **end for**
  $T_{FGP^B} \leftarrow max(T_{runs})$
  $T_{runs} = \emptyset$
  {Calculate time for Long Tail phase}
  $startState \leftarrow cutoffState$
  **while** $startState < stopState$ **do**
    $nextState \leftarrow startState + 1$
    **for** $i = 0$ **to** $NumRuns$ **do**
      $p \leftarrow startState$
      Randomly set $p$ elements of matrix $D^{ALL}(t)$ to 1
      **if** $STATIC\_CP\_CASE$ **then**
        $maxTime \leftarrow getMaxTimeStatic(startState, nextState, D^{ALL}(t), P^C)$ {For static contact patterns}
      **else**
        $maxTime \leftarrow getMaxTimeTV(startState, nextState, D^{ALL}(t), P_{TV}^C, MaxStepsDay)$ {For time-varying contact patterns}
      **end if**
      $T_{runs} \leftarrow T_{runs} \cup maxTime$
      **if** $maxTime = 0$ **then**
        $nextState \leftarrow nextState + 1$
      **else**
        $startState \leftarrow nextState$
        $nextState \leftarrow nextState + 1$
      **end if**
    **end for**
    $T_{LTP^B} \leftarrow T_{LTP^B} + max(T_{runs})$
  **end while**
  $T_{dss^B}^{upper} \leftarrow T_{FGP^B} + T_{LTP^B}$

$startState$ as $S(0)$ and $cutoffState$ as $S(\alpha*F)$ and predicts ($T_{FGP^B}$) to directly reach $cutoffState$ for several runs (i.e. $NumRuns$ using $getMaxTime$ method described in Algorithm 4 and Algorithm 6 for both static and time-varying contact patterns. Afterward, it starts Long Tail phase by initializing $startState$ to $S(\alpha*F)$ and $nextstate$ to $S(\alpha*F+1)$ and predicts maximum time spent in $nextstate$ using Algorithm 4 and Algorithm 6. If the model cannot reach $nextstate$ due to non-existence of *UUC* and *UDSC* processes then, $nextstate$ will be marked as non-reaching and $maxTime$ to reach this state will be set to 0. In this case, the model will again compute $maxTime$ by setting $nextstate$ to $S(\alpha*F+2)$ while keeping the same $startState$ as $S(\alpha*F)$. On the other side, if model could reach $S(\alpha*F+1)$ (i.e $maxTime \neq 0$) then, it sets $startState$ as $S(\alpha*F+1)$ and $nextstate$ as $S(\alpha*F+2)$. Likewise, the model repeats the same process until it reaches final state $S(F)$. Finally, the model computes $T_{LTP^B}$ as the sum of maximum time spent in each reachable state. To get better prediction of $T_{dss^B}^{upper}$, I run Fast Growing and Long Tail phase for $NumRuns$.

In this Section, I presented the detailed description of different approaches to predict the tighter upper bounds of data dissemination time and also presented how my Markov Model employs the Cut-off point approach to predict $T_{dss^B}^{upper}$ by dynamically estimating Cut-off point $\alpha$ using the Cut-off Estimator of INDIGO framework. In next Sections, I will present the methodology to create real data dissemination time from contact traces and also present the results obtained from different real-world contact traces using different approaches in Section 5.6.

## 5.5 Measured data dissemination time $T_{dss^B}^{meas}$ using real-world contact traces under broadcast data dissemination strategy

To compare the performance of upper bound of data dissemination time $T_{dss^B}^{upper}$ obtained from INDIGO framework, I simulate the real contact traces and measure the actual data dissemination time $T_{dss^B}^{meas}$ for each contact trace. Using these contact traces, I first fill in missing contacts to avoid any inaccurate measurement of $T_{dss^B}^{meas}$. For example, due to Bluetooth scanning interval, a device can discover neighbors only during the Bluetooth scan however, there could be a possibility that people were still in contact with each other during two consecutive scans. Finally, I prune the contact traces for isolated users (who do not have any contact with others).

From the final traces, I randomly select $M$ users as mobile data sources and

the remaining $N$ mobile users as those who are interested in gathering all data messages from all selected data sources. I also define a data matrix $D^{ALL}(t)$ and that store data messages collected by mobile users from different mobile users and data sources. At each time step, I update matrix $D^{ALL}(t)$ based on whether a contact happens between two users or a user and data source under broadcast data dissemination strategy (using *UUC* or *UDSC* described in Section 5.3). I stop the simulation at time $T_{dss^B}^{meas}$ when all elements of $D^{ALL}(t)$ become 1 or the contact traces is terminated. During these simulations of real-world contact traces, I also record the maximum possible fraction $DF_{max}$ of data that could be collected in the network by all nodes. To ensure comparable results, I also set maximum fraction of data collected $DF_{max}$ and measured data dissemination time $T_{dss}^{meas}$ through emulation of real traces as ground truth for the Markov Model of INDIGO under broadcast dissemination strategy. Please note that the $T_{dss^B}^{meas}$ will consist all real aspects of contact patterns i.e. time-varying contact patterns and multiple simultaneous contacts. Likewise, I create the ground truth data dissemination time $T_{dss^B}^{meas}$ and will utilize as a benchmark for both static and time-varying data dissemination modeling.

To capture the impact of diverse environments and to investigate the accuracy of INDIGO, I simulate 5 real-world contact traces (the Infocom 2005 trace (INFOCOM) Scott et al. [2006c], the PerCom 2012 trace (PERCOM) SCAMPI [2012], the Reality Mining trace (MIT) Eagle et al. [2009] , Crowded urban area trace collected at roller tour in Paris (ROLLERNET Tournoux et al. [2009]) and, MACACO traces MACACO [2012]) discussed in Chapter 3.

Further, to understand the impact of contact patterns during different time intervals and, to show the applicability of my model under different contact patterns, I use segments of contact traces that exhibit opposite behaviors. For conference environment INFOCOM trace, I take $2^{nd}$ day data when people are most active and shows high dynamics in their contacts, however for PERCOM trace I consider the day with less activity ($4^{th}$ day). Through ROLLERNET trace collected from a crowded urban area, I also investigate the impact of short span data (only for three hours) on the prediction of $T_{dss^B}^{upper}$. For university environment MIT trace, I validate my model for one-month data. Still, in this month, there is some variance among the weeks, so I use each week contact trace separately and call them as MIT-W1, MIT-W2, MIT-W3 and MIT-W4 traces. Finally, for our own collected MACACO traces collected in different countries and groups, I also consider the month where a maximum number of users make regular and intense use of the wireless network. The most active month for France and Brazil group were May 2015 and October 2015 respectively. Out of the selected months for France and Brazil and to understand the impact of contact patterns variance

Table 5.2. Simulation settings and $T_{dss^B}^{meas}$ for maximum data fraction for all weeks under broadcast data dissemination strategy.

| **Trace** | # **Mobile Users** | #**Data Sources** | $DF_{max}$ | $T_{dss^B}^{meas}$ (in secs) |
|---|---|---|---|---|
| INFOCOM ($2^{nd}$ day) | 30 | 9 | 1 | 65251 |
| PERCOM ($4^{th}$ day) | 32 | 9 | 1 | 32420 |
| ROLLERNET | 50 | 12 | 1 | 1554 |
| MIT-W1 (from $10^{th}$ month) | 51 | 15 | 0.9633 | 581440 |
| MIT-W2 (from $10^{th}$ month) | 57 | 15 | 0.9427 | 604748 |
| MIT-W3 (from $10^{th}$ month) | 61 | 15 | 0.9738 | 596457 |
| MIT-W4 (from $10^{th}$ month) | 64 | 15 | 0.9687 | 605577 |
| MACACO–France–$W^H$ | 15 | 4 | 0.9048 | 566403 |
| MACACO–France–$W^L$ | 15 | 4 | 0.8095 | 542396 |
| MACACO–Brazil–$W^H$ | 6 | 4 | 0.80 | 361425 |
| MACACO–Brazil–$W^L$ | 6 | 4 | 0.78 | 340460 |

among different weeks similar to MIT traces, I further segment one-month data into four weeks and only consider two weeks that exhibit opposite behavior in terms of contact patterns. For each group, I take the week having the highest number of contacts and the week having the lowest number of contacts and call them MACACO–France–$W^H$, MACACO–France–$W^L$, MACACO–Brazil–$W^H$ and MACACO–Brazil–$W^L$ for France and Brazil groups respectively. Table 5.2 presents the simulation settings and $T_{dss^B}^{meas}$ for maximum data fraction $DF_{max}$ collected in all segments of all traces. I will also use the similar settings in the Markov Model to predict $T_{dss^B}^{upper}$ through INDIGO framework.

In this Section, I basically presented how did I create the ground truth for real data dissemination time that can be utilized as a benchmark to predict the upper bound of data dissemination time for both static and time-varying contact patterns. In next Section, I will present and discuss the results obtained with my framework for both contact patterns under broadcast dissemination strategy.

## 5.6 Results and Discussion

In this section, I validate the upper bound of data dissemination time $T_{dss^B}^{upper}$ obtained from my Markov Model of INDIGO framework against the ground truth data dissemination time, $T_{dss^B}^{meas}$ from real-world traces. For each selected contact trace segment of each contact trace, INDIGO predicts $T_{dss^B}^{upper}$ for the maximum fraction of data collected $DF^{max}$ under both contact patterns and different approaches described in the previous section i.e. *Basic Data Dissemination Algorithm*, *Bisection*, and *Cut-off Point* approach. For all contact trace segments, the contact probabilities were predicted through the *Contact Probability Prediction Module* and the Cut-off point $\alpha$ was estimated with the help of Markov Model and Cut-off Estimator sub-components of *Data Dissemination Prediction Module*. Each experiment is repeated 500 times for statistical convergence. Now I will present the results for $T_{dss^B}^{upper}$ for both static and time-varying contact probabilities.

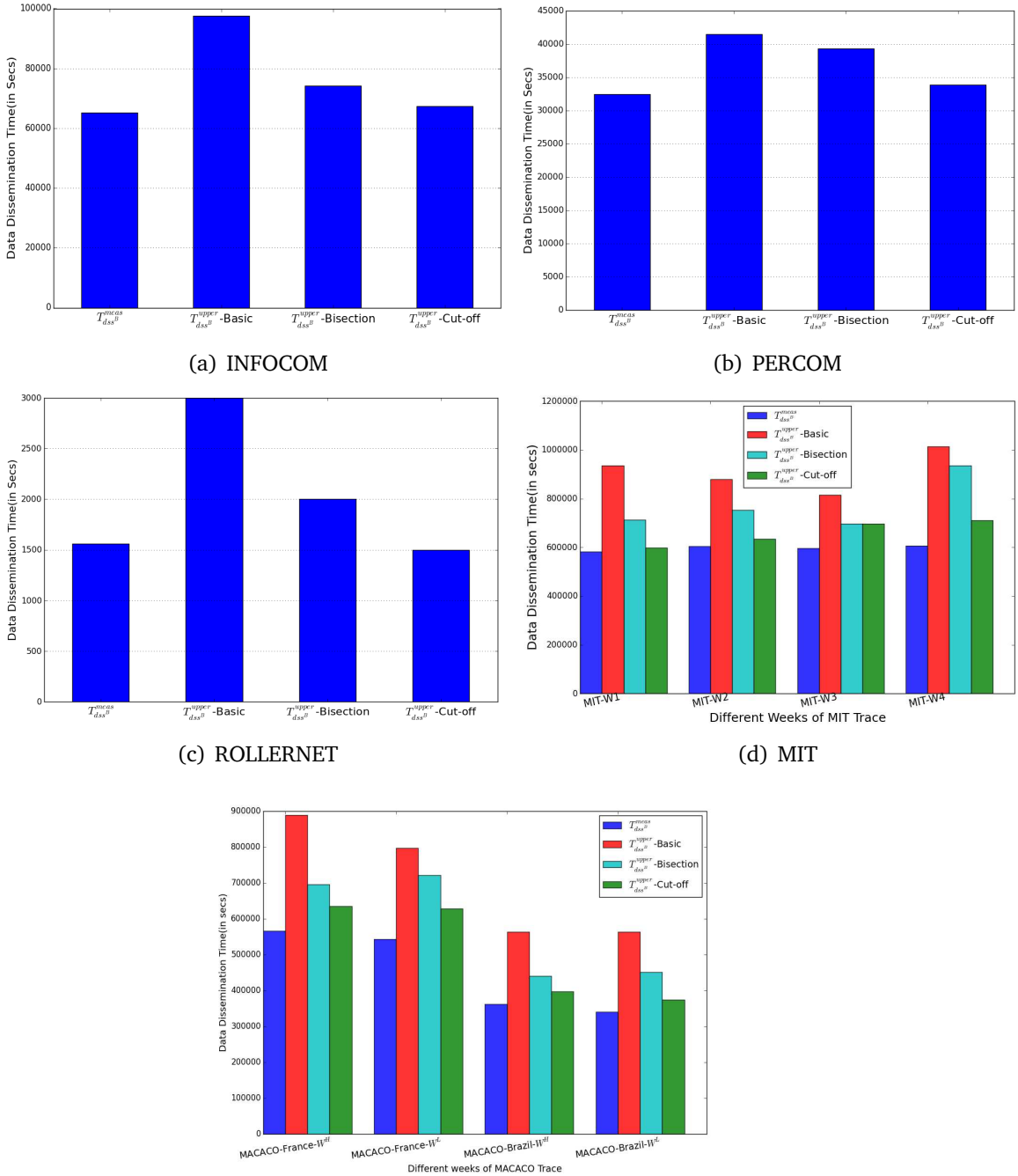
### 5.6.1 Static contact probabilities case

Figure 5.8 presents how well INDIGO is able to predict $T_{dss^B}^{upper}$ against $T_{dss^B}^{meas}$. Table 5.3 presents the Cut-off points utilized in the Cut-off point approach estimated through the Cut-off Estimator.

From the results, we observe that the Cut-off point approach achieves the tightest upper bound of data dissemination time $T_{dss^B}^{upper}$ against $T_{dss^B}^{meas}$. However, we can also see a gradual improvement in the accuracy of $T_{dss^B}^{upper}$ in Bisection approach as compared to the $T_{dss^B}^{upper}$ obtained through worst-case or basic data dissemination upper bound from Algorithm 3.

For INFOCOM trace, the Cut-off point approach of INDIGO achieves the tightest upper bound with 3.27% error compared to $T_{dss^B}^{meas}$. While for Algorithm 3 and Bisection approach, we observe an error of 49.51% and 13.84% respectively. This happens because $2^{nd}$ day of INFOCOM trace exhibits a high probability of contact among mobile users that significantly impacts the prediction of data dissemination time for Algorithm 3. Due to high contact rates, the data gets disseminated quickly in a real world scenario, while Algorithm 3 overestimates $T_{dss^B}^{upper}$ due to it's nature of calculating the *maximum time* for each possible state transition.

Similarly for another conference environment i.e. PERCOM trace, the Cut-off point approach outperforms and achieves the tightest upper bound of data dissemination time $T_{dss^B}^{upper}$ (with 4.38% an error against $T_{dss^B}^{meas}$). We can also observe that both Algorithm 3 (with an error of 27.51%) and Bisection approach (with an error of 21.2%) performs similarly. Compared to INFOCOM trace, since




(a) INFOCOM (b) PERCOM

(c) ROLLERNET (d) MIT

(e) MACACO

Figure 5.8. Comparison of real data dissemination time $T_{dss^B}^{meas}$ against the upper bounds of data dissemination time $T_{dss^B}^{upper}$ predicted using different approaches for INFOCOM, PERCOM, ROLLERNET, MIT and MACACO traces under broadcast dissemination strategy and static contact probabilities.

the contact rate among people is much lower in PERCOM trace, the rate of data gathering in the real world is also slow. Therefore, the maximum time calculated for each state in Algorithm 3 has comparatively less error. Due to lesser contact probabilities among people, the performance of Bisection approach deteriorates in PERCOM trace (compared to INFOCOM trace).

For ROLLERNET trace (in Figure 5.8(c)) from the crowded urban environment, we can also see that Cut-off based approach provides the tightest upper bound of data dissemination time $T_{dss^B}^{upper}$ with 10% error. However, we can observe that Algorithm 3 and Bisection approach highly overestimates data dissemination time as compared to $T_{dss^B}^{meas}$. This happens because in ROLLERNET traces, people were accumulated for a tour and contact probabilities among people are very high thus, data quickly get disseminated among people. However, Algorithm 3 overestimates it as it checks maximum time spent in each state. The data dissemination time calculated by Bisection approach is also high because it calculates maximum time in last 50% of the states.

Figure 5.8(d) depicts the results obtained from MIT traces for all 4 weeks (W1 to W4). We can observe that using Cut-off based approach, INDIGO is able to cope up with long duration contact traces and once again provides tightest upper bound. We can also see that for the first 3 weeks (MIT-W1 to MIT-W3), the upper bound $T_{dss^B}^{upper}$ is almost similar to the contact pattern of people are relatively identical. However, for the $4^{th}$ week, due to fewer contacts among people (semester break at MIT), I obtain much higher $T_{dss^B}^{meas}$ and also the higher error in upper bound estimation. In MIT traces, we can also notice that $T_{dss^B}^{upper}$ is comparatively less tight than INFOCOM and PERCOM traces (error from $T_{dss^B}^{meas}$ lies between 2% to 10%).

Finally, in Figure 5.8(e), I present the applicability of the INDIGO under broadcast and static time-varying contact probabilities for our collected MACACO traces. My results show that the Cut-off approach once again provides the tighter prediction of $T_{dss^B}^{upper}$ for both France and Brazil traces and for both weeks. For all weeks of both groups, we can observe that INDIGO predicts $T_{dss^B}^{upper}$ within 9-15% error for Cut-off point approach. Further, we can also see that Algorithm 3 and Bisection approach overestimates $T_{dss^B}^{upper}$ similar to MIT traces. More specifically, we see that for France-$W^H$, for Algorithm 3 upper bounds are much looser as compared to France-$W^L$. This happens because, in reality, the contact probability among people is higher while Algorithm 3 estimates maximum time each step, therefore, the overall time predicted to get much higher as compared to real data dissemination time.

From the above discussion, I conclude that my *Cut-off point* approach outperforms in all traces for static contact patterns under broadcast data dissemination

Table 5.3. Cut-off points estimated through `Cut-off Estimator` for all traces for Cut-off point based approach.

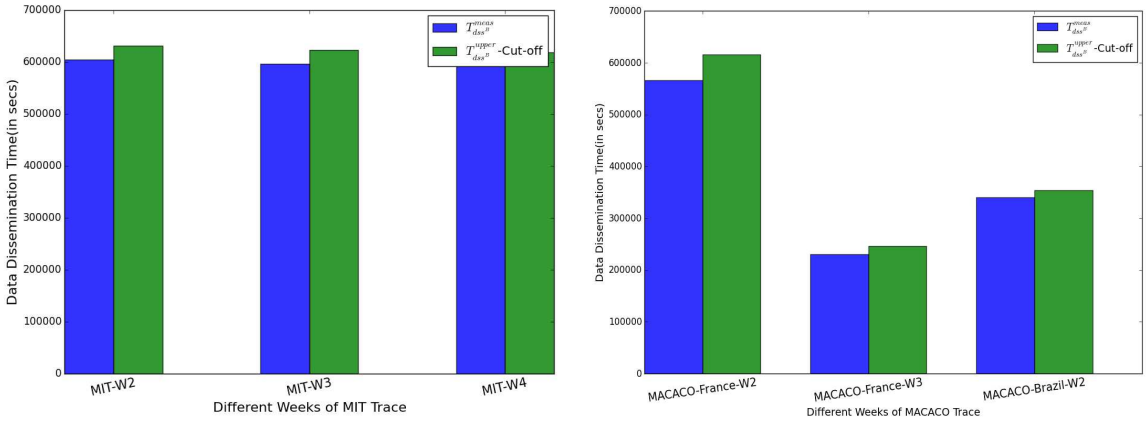
| Trace | Cut-off Point |
|---|---|
| **INFOCOM** | 0.70 |
| **PERCOM** | 0.65 |
| **ROLLERNET** | 0.96 |
| **MIT-W1** | 0.80 |
| **MIT-W2** | 0.85 |
| **MIT-W3** | 0.961 |
| **MIT-W4** | 0.85 |
| **MACACO–France–**$W^H$ | 0.9048 |
| **MACACO–France–**$W^L$ | 0.70 |
| **MACACO–Brazil–**$W^H$ | 0.62 |
| **MACACO–Brazil–**$W^L$ | 0.75 |

strategy. Therefore, we can say that the utilization of $\alpha$ significantly improves the upper bound of data dissemination time $T_{dss^B}^{upper}$ and outperforms as compared to the *Bisection* and *Basic Data Dissemination* algorithm. Therefore, in further Sections, I will only present the results obtained from the *Cut-off approach* based on its suitability for the tighter prediction of $T_{dss^B}^{upper}$.

### 5.6.2 Time-varying contact probabilities case

In this Section, I will present the upper bound of data dissemination time $T_{dss^B}^{upper}$ under the time-varying contact patterns and broadcast data dissemination strategy using the Cut-off approach. I present the prediction results for those contact traces those are collected for a longer duration of time and exhibits certain regularities in the contact patterns among people (detailed description is in Chapter 4). Therefore, for time-varying contact probabilities, I present the results predicted for MIT and MACACO traces.

Figure 5.9 presents the applicability of INDIGO in predicting $T_{dss^B}^{upper}$ against $T_{dss^B}^{meas}$ for the different weeks of MIT and MACACO traces. For MIT trace, the *Contact Probability Prediction Module* of INDIGO trains the contact probability pre-

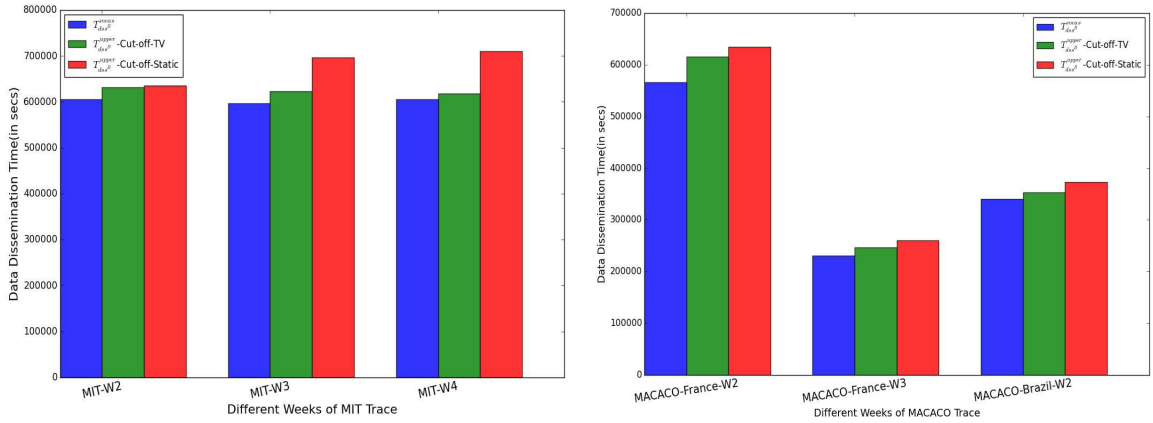



(a) MIT (b) MACACO

Figure 5.9. Comparison of real data dissemination time $T_{dss^B}^{meas}$ against the upper bounds of data dissemination time $T_{dss^B}^{upper}$ predicted using Cut-off approach for different weeks of MIT and MACACO traces under broadcast dissemination strategy and time-varying contact patterns.

diction model with MIT-W1 contact probabilities that further predicts the contact probabilities for each pair of people for rest of the weeks. Once the contact probabilities for other weeks of MIT trace are predicted and given as an input to the *DDT-Markov* then INDIGO predicts $T_{dss^B}^{upper}$ for all these weeks and compare it with $T_{dss^B}^{upper}$. Similarly, for MACACO trace, the *Contact Probability Prediction Module* trains the model for the first 1-week trace of both France and Brazil and predicts the contact probabilities for future weeks. For the contact trace of France, there were only three weeks (W1, W2, and W3) where I had enough contacts of people, therefore, I present $T_{dss^B}^{upper}$ values for W2 and W3. Similarly for Brazil trace, the enough contacts were available only for 2 weeks of data, therefore, I present my prediction result for W2 of Brazil traces and use W1 for learning.

From Figure 5.9, we can observe that employing time-varying contact probabilities also provides tighter upper bound of data dissemination time and we can also observe that error in prediction of $T_{dss^B}^{upper}$ decreases for subsequent weeks i.e. prediction error in $MIT-W4 < MIT-W3 < MIT-W2$. It happens because of more training data provided to the contact probability estimator model thus leads to better prediction of contact probabilities. We can also observe the similar trend in MACACO traces. Further, during the validation of time-varying contact probabilities on the prediction of $T_{dss^B}^{upper}$, we can further observe that estimation of Cut-off point $\alpha$ also changes from the $\alpha$ of Cut-off approach employed in static contact patterns case. The change in $\alpha$ is reasonable due to its dynamic estima-




(a) MIT (b) MACACO

Figure 5.10. Comparison of real data dissemination time $T_{dss^B}^{meas}$ against the upper bounds of data dissemination time, $T_{dss^B}^{upper}$ predicted using Cut-off approach for different weeks of MIT and MACACO traces under broadcast dissemination strategy and both static and time-varying contact patterns.

tion obtained from Cut-off Estimator. It happens because the estimation of $\alpha$ is directly proportional to the rate of data gathering (or fraction of data collected) that is driven by contact patterns of people.

Further, in Figure 5.10, I altogether present the prediction of $T_{dss^B}^{upper}$ against $T_{dss^B}^{meas}$ obtained for both static and time-varying contact probability under broadcast data dissemination strategy. The Figure clearly shows that utilization of time-varying contact patterns provides much tighter upper bound of data dissemination time as compared to the static contact patterns. Therefore, with my approach of time-varying contact patterns prediction and Cut-off point based approach I achieved realistic and tighter upper bound of data dissemination time for real-world contact traces against the ground truth $T_{dss^B}^{meas}$ .

Finally, I also compare the results obtained from INDIGO with the state of the art key contributions for static contact patterns over time under the broadcast strategy. For time-varying contact patterns, there is not any significant work that is comparable to my model and traces. For static contact patterns, the key contributions I considered are Picu et al. [2012], Shah [2009] and Boldrini et al. [2014]. The work that is closely related to my thesis is from Picu et al. [2012] where the authors present an approach to predict the upper bound of data dissemination time using the conductance property of contact graphs by considering heterogeneous mobility. However, the author only focuses on single source and sequential contact models i.e. they do not consider the impact of multi-contact

Table 5.4. Error obtained in predicting $T^{meas}_{dss^B}$ from state of the art work as compare to our approach.

| **SOA Methods** | **Error for INFOCOM Trace (in %)** | **Error for MIT Trace (in %)** |
|---|---|---|
| Shah et. al (2009) | 200 | 500 |
| Picu et. al (2012) | 15 | 50 |
| Boldrini et. al (2014) | 50 | NA |
| INDIGO | 4 | 10 |

and multi-source data dissemination on the upper bound of data dissemination time. For all relevant state of the art works, I present the average error obtained from their approaches while predicting the upper bound of data dissemination time $T^{upper}_{dss^B}$ against the real simulation time $T^{meas}_{dss^B}$ in Table 5.4 and find that INDIGO framework performs *much* better than the existing approaches also for static contact patterns case.

## 5.7 Conclusions

In this Chapter, I presented different approaches to predict the upper bound of data dissemination time using the *Broadcast Sub-Module* of INDIGO framework under broadcast data dissemination strategy for both static and time-varying contact patterns. I presented how the `Markov Model` employs different approaches to predict the upper bound of data dissemination time by modeling multi-source multi-contact data dissemination process as a Markov chain and predicts the upper bound of data dissemination time by predicting the maximum time spent in each state utilizing the contact probabilities among people (either static or time-varying). I started with the *Basic Data Dissemination Algorithm* that sums up all maximum time predicted in transiting from each state of the Markov chain. However, the results obtained from the *Basic Data Dissemination Algorithm* did not serve the purpose to provide tighter bounds of data dissemination time. Therefore, I further observed the real-world data gathering process and utilize the exponential cut-off property of inter-contact time distribution that impacts the data gathering process to provide tighter upper bounds of data dissemination time. To do this, I divided the data gathering process into two parts i.e. *Fast Growing* and *Long Tail* phase and introduced $\alpha$ Cut-off point that can be dy-

namically estimated between the communication of `Markov Model` and `Cut-off Estimator`. Finally, I also presented the modified algorithm (Algorithm 7) that utilizes Cut-off point $\alpha$ to predict the upper bound of data dissemination time for both static and time-varying contact patterns.

Later on, I presented the upper bounds of data dissemination time obtained for all real-world contact traces for static contact patterns and for MIT and MACACO traces with time-varying contact patterns. The reason behind considering only MIT and MACACO for time-varying contact pattern is because learning of contact probabilities require the longer duration of contact data that exhibits some regular patterns. The results for both static and time-varying contact probabilities validate INDIGO framework due to the tighter upper bounds of data dissemination time predicted from it. Finally, I also showed that consideration of time-varying contact probabilities further tightens the upper bound of data dissemination time as they reflect realistic mobility patterns of people. The methodology and results presented in this Chapter provide a complete modeling that collectively considers different aspects of mobility contact patterns and can help in providing realistic and tighter upper bound of data dissemination time under Broadcast Strategy for Type II case of INDIGO framework. The key findings of this Chapter are:

- To the best of my knowledge, INDIGO is the first work that predicts tighter upper bound of data dissemination time by collectively considering real world mobility and communication aspects under the broadcast strategy.
- I identified the long tail Cut-off behavior in data dissemination process and also proposed a Cut-off approach that predicts the tighter upper bound of data dissemination time. To the best of my knowledge, this observation has not been identified before. Such discovery is very important as it has a strong impact on the dissemination time.
- For static contact patterns, the proposed Cut-off point based approach performs much better as compared to the existing works in literature.
- The Cut-off approach further tightens the prediction using time-varying contact patterns. Prediction from time-varying contact patterns improves with time due to more training data.

The work done in this Chapter has resulted in 3 publications at IEEE INFOCOM 2013, ACM HP-MOSys 2013 co-located with MSWim 2013 and IEEE Med-Hoc-Net 2015 conferences. In next Chapter, I will consider the interest-driven data dissemination strategy while predicting the upper bound of data dissemination time for Type III case and also present the methodology to learn interests of

people and similarity among them. I will also present the second component of the *Data Dissemination Prediction Module* i.e. *Interest-Driven Sub-Module* in detail.

# Chapter 6

# Prediction of Upper Bound of Data Dissemination Time Under Interest-driven Strategy

## 6.1 Introduction

In the previous Chapter, I presented the prediction of the tighter upper bound of data dissemination time under the broadcast strategy for both static and time-varying contact patterns. From the results obtained from INDIGO for broadcast strategy, I found Cut-off based approach as the best performing.

In this Chapter, I will focus on a more realistic aspect of data dissemination process i.e. interest-driven data dissemination strategy where people collect and share information that is interesting to them as opposed to the broadcast approach that enforces people to receive all of the information. Along with interest-driven data dissemination strategy, I also take into account the heterogeneous contact patterns, multiple simultaneous contacts among people and data originating from multiple data sources. The importance of interest-driven data dissemination modeling along with heterogeneous mobility is also highlighted in literature Mei et al. [2011] Ciobanu et al. [2015]. However, these works do not focus on real interests captured along with user mobility rather they rely on a publish-subscribe scheme where users explicitly show their interests in certain topics. Asking user interests are neither feasible in long term nor scalable because it limits the wide range of interests a user can express and their validity over a long time. Thus, learning of user interests is a more practical approach to model interest-driven data dissemination. The interest learning part of this Chapter is done in collaboration with Telefònica Research during my internship.

This Chapter addresses the different aspects discussed above and presents my approach to predict the tighter upper bound of data dissemination time under interest-driven data dissemination strategy by collectively considering different aspects (heterogeneous mobility patterns both static and time-varying, multiple simultaneous contacts and interest-driven data dissemination strategy) of data dissemination using the `Interest-Driven Sub-Module` of INDIGO framework. My approach can be utilized for both Type III case of the Physical–Social proximity table where we have both physical and social proximity information. The physical proximity information is derived from contact patterns while the interest similarities among people represent their social proximity.The `Interest-Driven Sub-Module` utilizes Markov-chain based prediction model that employs the Cut-off approach using predicted pair-wise contact probabilities and enables automatic learning of web interests of people for tighter upper bounds prediction of data dissemination time. To the best of my knowledge, INDIGO is the first work that predicts tighter upper bound of data dissemination time by learning real interests of people along with heterogeneous mobility aspects.

The chapter is structured as follows. In Section 6.2, I will present the detailed description of the components required to predict the upper bound of data dissemination time under interest-driven data dissemination strategy. Further, in Section 6.3, I will present how do we learn the interests of people from their web browsing history by employing information retrieval techniques. Section 6.4 presents the modified Cut-off approach for interest-driven data dissemination strategy for static and time-varying contact patterns. Afterward, in Section 6.5, I present the methodology to create ground truth data dissemination time obtained from different contact traces for interest-driven data dissemination strategy. Section 6.6 presents results obtained from INDIGO for contact traces of diverse environment and duration. Finally, I conclude the Chapter with Section 6.7.

## 6.2 Overview of INDIGO framework components required under interest-driven strategy

Figure 6.1 outlines the different components of `Interest-Driven Sub-Module` of INDIGO under interest-driven data dissemination strategy. To predict the upper bound of data dissemination under interest-driven data dissemination strategy, INDIGO requires the *Contact Probability Prediction Module*, *DDT-Markov Component* and *Interest Learning Component*. The pair-wise heterogeneous contact

patterns are predicted using the *Contact Probability Prediction* module for both static and time-varying contact patterns using the real world or synthetic traces (details are in Chapter 4). INDIGO also learns real interests of people with the help of *Interest Learning Component* using their on-mobile web browsing history and also calculates pair-wise interest similarities between people. Finally, predicted contact probabilities and interest similarities are given as an input to the *DDT-Markov Component* to predict the upper bound of data dissemination time utilizing the Markov chain based model. Table 6.1 presents the additional mathematical notations used for interest-driven data dissemination strategy (for other common notations please refer Table 5.1 of Chapter 5).

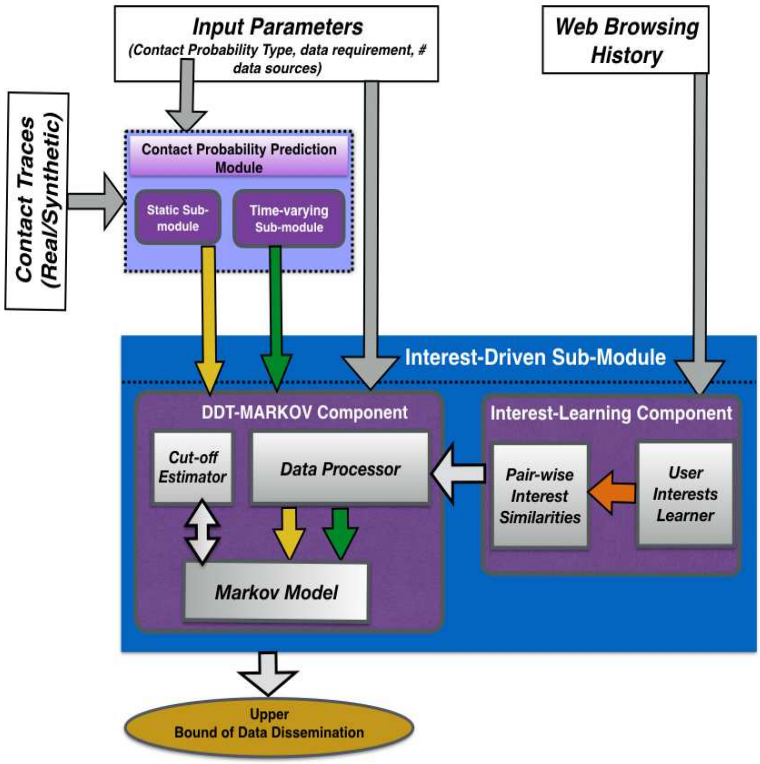



Figure 6.1. Different components of INDIGO required to predict upper bound of data dissemination time under interest-driven data dissemination strategy.

### 6.2.1 Contact Probability Prediction Module

The working of *Contact Probability Prediction Module* and its utility in prediction of data dissemination time is already described in detail in the previous Chapters 4, 5. Similar to broadcast data dissemination strategy it provides contact

Table 6.1. Notations used in INDIGO for Type III case.

| Notations | Description |
|---|---|
| $sim_{ij}$ | interest cosine similarity between any pair $i, j \in V$, $sim_{ij} \in [0,1]$ |
| $I^S$ | interest similarity matrix of all $sim_{ij}$ |
| $\mathbb{E}(I^S)$ | expected value of $I^S$ |
| $\beta$ | interest similarity threshold, $\beta \in [0,1]$ |
| $ut_{ij}$ | utility value (0 or 1) of data exchange between any pair $i, j \in V$ |
| $\mathbb{UT}$ | utility matrix of all utility values $ut_{ij}$ to determine data exchange based on interests |
| $T_{FGP^I}$ | maximum time spent Fast Growing Phase |
| $T_{LTP^I}$ | maximum time spent Long Tail Phase |
| $T_{dss^I}^{upper}$ | predicted upper bound of data dissemination under interest-driven data dissemination strategy |
| $T_{dss^I}^{meas}$ | measured data dissemination time from real traces under interest-driven data dissemination strategy |

probabilities for each pair of people $i$ and $j$ and *Contact Probability Type* parameter (0 for static contact probabilities and 1 for time-varying contact probabilities). For static contact patterns (i.e. *Contact Probability Type* $= 0$) it constructs a contact probability matrix $P_S^C$ with static pair-wise heterogeneous contact probabilities and for time-varying contact patterns (i.e. *Contact Probability Type* $= 1$), it constructs a contact probability matrix set $P_{TV}^C$ that contains pair-wise heterogeneous contact probabilities for each pair in different days.

### 6.2.2 Interest-Driven Sub-Module

This sub-module come under the *Data Dissemination Prediction Module* of INDIGO framework for Type III case (from Figure 3.2 of Chapter 3) and is responsible for predicting the tighter upper bound of data dissemination time for interest-driven dissemination strategy by learning the interests of people. It consists of `Interest Learning Component` to find interest similarities between people after learning their interest and *DDT-Markov Component* that further has `Data Processor`, `Markov Model`, and `Cut-off Estimator` sub-components. The `Data Processor` pre-processes different inputs required by the `Markov Model` which is the core of *DDT-Markov* and employs a Cut-off point based approach to predict tighter upper bound of data dissemination time for both static and time-varying contact patterns. The `Markov Model` communicates with `Cut-off Estimator`

to estimate the Cut-off point $\alpha$ that plays a significant role to provide the tighter prediction. The Cut-off point based approach incorporates realistic aspects of human mobility (i.e. heterogeneous contact patterns and multiple simultaneous contacts among people) and data dissemination strategy (i.e. interest-driven data dissemination as opposed to Broadcast strategy). Now, I will present the different components of *Interest-Driven Sub-Module* required to predict the upper bound of data dissemination time, $T_{dss^I}^{upper}$.

### Interests Learning Component

This component of INDIGO plays the most important role in learning interests of people and in providing the interest similarities between a different pair of people. People usually people prefer to receive and share information according to their interests (interest-driven) rather than receiving every possible information (broadcast). The `Interest Learning Component` enables interest-driven data dissemination by learning real web interests of people from their mobile browsing history. The semantic categories of visited websites like social networking, news, shopping etc. reflect user's web interests. In case, the web browsing history is not available in contact traces then, the `Interest Learning Component` synthetically creates the interests of people based on small, medium and weak ties concept of social networks using power-law distribution. For both cases, to determine the likelihood to exchange data among people, it calculates cosine similarity between extracted interests. I will discuss the `Interest Learning Component` in detail in next Section.

### Data Processor

This component processes all inputs coming from *Contact Probability Prediction Module*, *Interest-Learning Component* and *Input Parameters* for both static and time-varying contact patterns. The input parameter for interest-driven data dissemination strategy is the # of data sources and data requirement that allows INDIGO to predict data dissemination time bound for a different fraction of data. In the case of interest-driven strategy, it is more likely to collect a certain fraction of data as people will not always be interested to collect complete data fraction. Out of all users, it randomly assigns $M$ users as data sources and marks the rest as $N$ mobile users. Based on the type of contact probabilities (i.e. static or time-varying), it gives either $P_S^C$ contact probability matrix or $P_{TV}^C$ contact probability matrix set as an input to the `Markov Model`. Further, the interest similarity threshold $\beta$ is calculated as the Expected value of interest similarities matrix $I^S$

(Please note that $\beta$ can also be enforced as input to the model). Finally, based on $\beta$ and $I^S$, it prepares a Utility matrix $\mathbb{UT}$ as follows:

$$\mathbb{UT} = \{ut_{ij}\} \ \forall \ i, j \in V \tag{6.1}$$

$$ut_{ij} = \begin{cases} 1 & if \quad sim_{ij} \geq \beta \\ 0 & \text{otherwise} \end{cases}$$

$$\text{Where } \beta = \mathbb{E}(I^S)$$

Both $P^C_S$ or $P^C_{TV}$ and $\mathbb{UT}$ will be given as an input to `Markov Model` where, $\mathbb{UT}$ determines interest-driven data exchange while $P^C_S$ or $P^C_{TV}$ drives the mobility and heterogeneous contact patterns for the model. It is important to note that data message can only be exchanged if a pair of users has utility value as 1. This signifies the importance of interests similarities in real-world data dissemination process.

### Markov Model

This working of this sub-component is similar to the one used for broadcast strategy in Chapter 5. It is a Markov-chain based model that utilizes a Cut-off point based approach to mimic real-world data gathering process by collectively considering different real-world aspects of data dissemination. The `Markov Model` communicates with `Cut-off Estimator` to dynamically estimate the Cut-off point for the tighter prediction of the upper bound of data dissemination time (Please refer Chapter 5 for more details). Each state $S(x)$ of the `Markov Model` represents the total number of messages $(x)$ collected by mobile users from different data sources and the transition from one state to another state is driven by the mobility and contact patterns of users, their interest similarities and data sources. Once all nodes fulfill the data requirement then the `Markov Model` stops transiting and remain in the same state (absorbing state).

### Cut-off Estimator

This sub-component contributes towards the tighter prediction of the upper bound of data dissemination time by dynamically providing $\alpha$ value to `Markov Model`. Similar to broadcast strategy, it helps in calculating $\alpha$ by communicating the fraction of data collected in one state transition to `Cut-off Estimator` from `Markov Model`. Based on the fraction of data collected in one state the `Cut-off`

`Estimator` measures the change in data fraction and repeat this process until it reaches to a data fraction $\alpha$ beyond which change in data fraction becomes smaller than $\epsilon$. The calculated $\alpha$ value is communicated back to `Markov Model` using Equation 5.1 described in Chapter 5.

## 6.3 Learning of interests for interest-driven data dissemination

To model interest-driven data dissemination process, we need to understand the interests of people and also their willingness to share information with each other. The work done in literature have not considered real interests of people along with their heterogeneous mobility rather they rely on a publish-subscribe scheme where users explicitly show their interests in certain topics. Asking user interests are subjective and not scalable in long term as interests of people changes over time Boldrini et al. [2008] Mei et al. [2011] Ciobanu et al. [2015]. Thus, a user interests learning is more practical and effective approach to model interest-driven data dissemination.

To address this problem, I provide a solution to learn the interests of people through their browsing history by extracting the semantic categories of the websites using `Interest-Learning Component` of INDIGO. The semantic categories of visited websites like social networking, news, shopping etc. reflect user web interests. Different works on online advertisements suggest that user profiles built from website categories are an efficient method for user interest profiling Carrascosa et al. [2014]. Our own collected MACACO traces captures interests of people (on-mobile browsing history) along with their real mobility patterns (Wi-Fi connectivity). To the best of my knowledge, this is also the first contact trace that collects both information simultaneously. Further, to ensure the privacy of volunteers, I anonymize their identity and only utilize website host to build user interests. I neither use complete URL of the website nor examine the HTML content of a webpage.

In the case of contact traces that do not capture web browsing history, the `Interest-Learning Component` synthetically creates the interests of people by utilizing the concept of strong, medium, and weak ties computed with the power law distribution. To do this, it loops over all possible pairs of people and randomly draws $K$ interest weights from a power law distribution. After learning/creating the interests of people the `Interest-Learning Component` determines the likelihood to exchange data among people by calculating the cosine

similarity of their web or synthetic interests weight vectors.

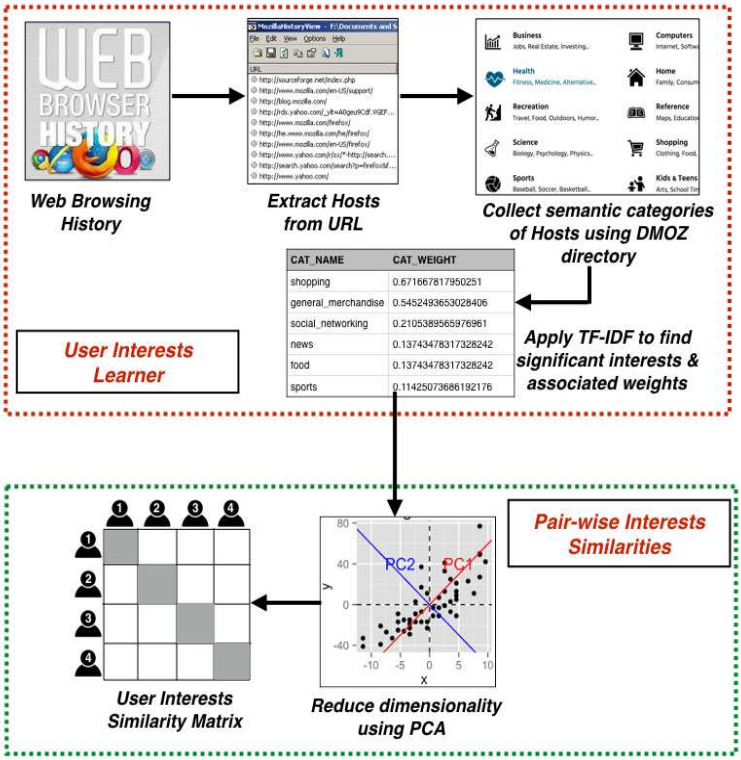



Figure 6.2. Working of Interest-Learning Component of INDIGO framework using User Interests Learner and Pair-wise Interests Similarities subcomponents from browsing history.

Figure 6.2 presents the working of *Interest-Learning Component* that contains two sub-components: `User Interests Learner` and `Pair-wise Interests Similarities`. The `User Interests Learner` investigates web browsing history from user's mobile and finds the hosts associated with each URL. Later, it queries DMOZ[1], a commonly used open directory of websites, to annotate the destination hostnames with semantic tags to obtain web categories of each host. The DMOZ directory returns a category of multiple hierarchies each host e.g., "Europe/News and Media" for host "bbc.com" or World/Region/Shopping" for host "amazon.com". I take into account the categories of all hierarchies for user's web interests construction. To learn user interests, I apply Term Frequency Inverse Document Frequency (TF-IDF) weighting scheme and create a term vector for each user that contains the web category and weight associated with this category Blei et al.

[1]http://www.dmoz.org/

[2003]. TF-IDF emphasizes on categories that distinguish a user from others by considering frequency count of each category (the TF term) and then scaling of frequencies for commonly used categories across all users (the IDF term). The IDF term effectively decreases the weight of categories that commonly appear across different users like higher hierarchy categories that are abstract.

After computing the most important web categories for each user, the `Pair-wise Interests Similarities` sub-component creates similarities among each pair of users in two steps:

1. I first apply Principal Component Analysis (PCA) on the interests vectors to reduce the dimensionality of these sparse interests term vectors that use singular vector decomposition to create a smaller set of dimensions. It's a popular technique that use singular vector decomposition to create a smaller set of dimensions by analyzing similarities between original dimensions.

2. From reduced interest term vectors, I prepare an Interest Similarity Matrix $I^S$ for all users using cosine similarity.

$$I^S = \{sim_{ij}\} \; \forall \; i, j \in V \tag{6.2}$$

Similarly, Figure 6.3 presents the working of *Interest-Learning Component* of INDIGO framework to create the synthetic interests of users using the power law distribution method. Once the interests of users are created then, the `Pair-wise Interests Similarities` computes the pair-wise cosine similarities and prepares the Interest Similarity Matrix $I^S$.

In order to verify the applicability of power law distribution for synthetic traces and observe the existence of strong, medium and weak ties concept in web interests, I plot user's real web interests learned from the MACACO trace and present the results for sample users in Figure 6.4 for both France and Brazil group. The plots shown in Figure 6.4 verifies the usefulness of power law distribution and the synthetic interests generated through the *Interest-Learning Component* of INDIGO framework.

I applied the above described approach to the different samples of web browsing history at Telefònica Research. For our contact traces, I only had the browsing history for our own collected MACACO traces while for other traces (INFOCOM, PERCOM, ROLLERNET, and MIT) unfortunately, we did not have the web browsing history, therefore, I rely on synthetic interests for such traces. Figure 6.5 presents the sample synthetic interests for MIT trace and real web interests generated from MACACO trace for both France and Brazil groups.

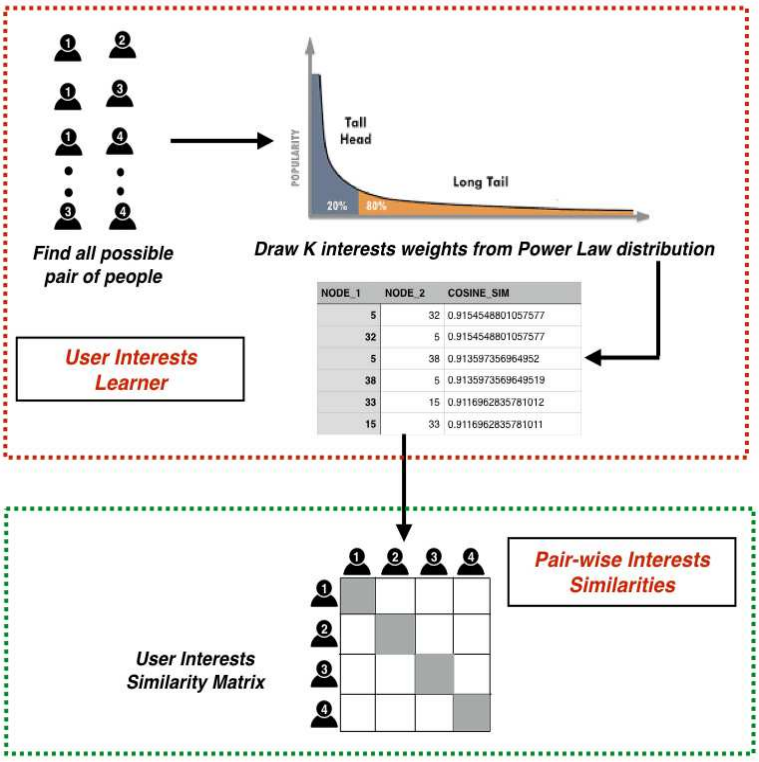
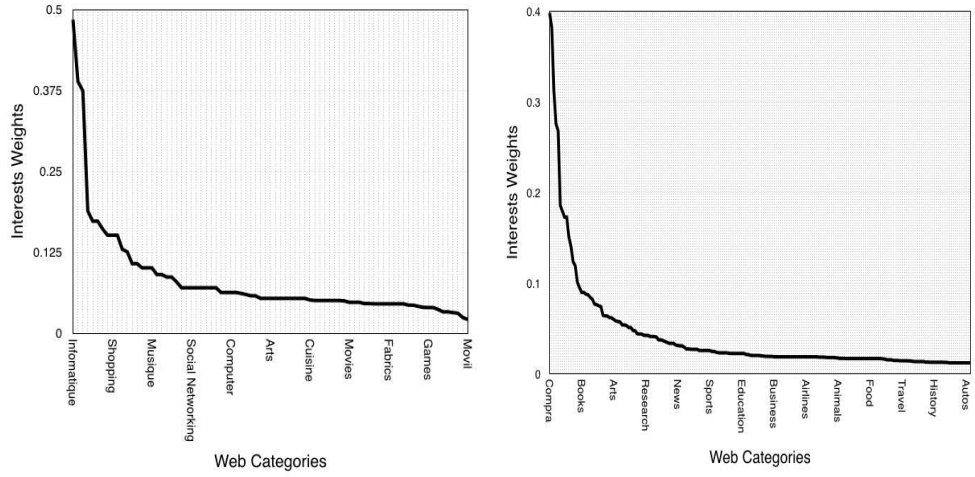



Figure 6.3. Working of Interest-Learning Component of INDIGO framework using User Interests Learner and Pair-wise Interests Similarities subcomponents for synthetic web interests using Power Law distribution.

From Figure 6.5, I observe that the *Interest-Learning Component* also produces some interests specific to country language. Further, Figure 6.6 presents interest similarities among volunteers in France and Brazil and shows that on average volunteers in Brazil group have higher interest similarities as compared to the France group. This could happen because, in France traces, volunteers range from students to staff of different departments and age groups while in Brazil traces most of the people were researchers from the same group.

In next Section, I will present how INDIGO incorporates the interests in the Cut-off approach to predict the upper bound of data dissemination time for both static and time-varying contact patterns. Please note that in this Chapter, I will only use the Cut-off based approach as it was proved as the best approach to predict the tighter upper bound of data dissemination time in the previous Chapter.




(a) MACACO–France (b) MACACO–Brazil

Figure 6.4. Distribution of sample user's web interests for MACACO France and Brazil groups. This distribution shows the applicability of synthetic web interests generated through INDIGO framework.

## 6.4 Tighter prediction of $T_{dss^I}^{upper}$ using Cut-off approach

Similar to the broadcast scheme, the `Interest-Driven Sub-Module` also models the multi-source and multi-contact data dissemination using the Markov chain based model under interest-driven data dissemination strategy. The Markov-chain based model utilizes the Cut-off point based approach to mimic real-world data gathering process. Each state $S(x)$ of our Markov-chain based model represents the total number of messages $(x)$ collected by mobile users from different data sources.

### 6.4.1 Preliminaries

Let us consider a network with $V = U \cup D$ users where $U = \{u_1, u_2, ....., u_N\}$ represents $N$ mobile users and $D = \{d_1, d_2, ....., d_M\}$ represents $M$ data sources (mobile or static). We assume that every data source $d_i \in D$ has a distinct data message $msg_i$. Further, the maximum number of messages gathered by a mobile user $u_j \in U$ is determined according to $\mathbb{D}$ obtained as an input parameter. By default, our approach predicts data dissemination bounds for different $\mathbb{D}$ from $\alpha$ to 1 within an interval of 0.05 ($\alpha$, $\alpha + 0.05$, $\alpha + 0.10$.....1). Please note that $\mathbb{D}$ represents the maximum fraction of data that needs to be collected by all mobile

| NODE_ID | INTEREST_WEIGHT_1 | INTEREST_WEIGHT_2 | INTEREST_WEIGHT_3 | INTEREST_WEIGHT_4 | INTEREST_WEIGHT_5 |
|---|---|---|---|---|---|
| 5 | 0.038677060418469944 | 0.05848175390225149 | 0.06293147030628435 | 0.024505179509199338 | 0.08662082782317443 |
| 6 | 0.07927557258827994 | 0.07784224803397985 | 0.10021105766667213 | 0.05643978028840625 | 0.04556246360546721 |
| 7 | 0.0908627724226839 | 0.04400937654483683 | 0.06939682735361281 | 0.09451284131851984 | 0.02448573777677102 |
| 8 | 0.03298094765910305 | 0.08182810500755161 | 0.07414812805795269 | 0.08369522341089614 | 0.03062999566915446 75 |
| 9 | 0.06474726517700599 | 0.00154975370326311986 | 0.02728593707482486 | 0.06534964760417708 | 0.03548244332135598 |
| 10 | 0.05706650145244549 | 0.07881560323776157 | 0.02734531816760403 3 | 0.02405107645128358 | 0.08515747232034782 |
| 11 | 0.02297982244800256 | 0.01655362777680005 34 | 0.08341821765281179 | 0.08407672383034027 | 0.04941760570137690 4 |
| 12 | 0.08463508060281623 | 0.06689043421154127 | 0.07024846771685854 | 0.06440973770059473 | 0.08395713336233292 |
| 15 | 0.02551734766817425 | 0.02886577076761447 5 | 0.00108891241395491 08 | 0.04979722902370861 | 0.08602357574562995 |
| 16 | 0.01791706282979977 | 0.089000033907817 5 | 0.07093162996876973 | 0.05364263489892233 | 0.04604964520677489 |
| 17 | 0.03710585348605315 | 0.08119998845548372 | 0.05630334448035164 | 0.01115206459050278 7 | 0.03649162504102404 |
| 19 | 0.09577552573378307 | 0.03013394769818524 7 | 0.04925175260167659 | 0.07232047666012797 3 | 0.01134593839202395 |
| 20 | 0.08610661899939603 | 0.02736601872915295 | 0.06596825641185886 | 0.07455679126005935 | 0.0709611397690863 |
| 21 | 0.07865194804875242 | 0.04551400994103109 5 | 0.07619056983281393 | 0.07377167269939646 | 0.02086956959732941 6 |
| 22 | 0.00854159168591322 5 | 0.10492277675467501 | 0.08915371576397475 | 0.03380870915237052 4 | 0.10455106899253508 |
| 23 | 0.06264530302273244 | 0.00718869372970954 05 | 0.09788197309672994 | 0.02831298873284181 | 0.04555431387720072 |
| 25 | 0.09578436072966138 | 0.05680022613448219 4 | 0.06967822322655055 | 0.01067684254283365 | 0.05019765807398988 5 |
| 26 | 0.08730051691194615 | 0.0290571255387042 | 0.06695649452594369 | 0.08526758098758078 | 0.0620085201842388 |
| 27 | 0.02491358360705129 8 | 0.03561725303447095 | 0.05219020802731088 | 0.04544063855141437 | 0.0293815013218528 |
| 28 | 0.01212982313273113 | 0.02945627230994129 | 0.02358838641736574 2 | 0.0681174870922598 | 0.04983354935083122 |
| 29 | 0.02415478519084924 | 0.01322373451598217 | 0.05708073307095376 | 0.04917858233741846 | 0.07631666983942523 |
| 30 | 0.03203174090516165 | 0.06215318037608474 | 0.04110297965111864 | 0.08509710752664792 | 0.09400013765223916 |

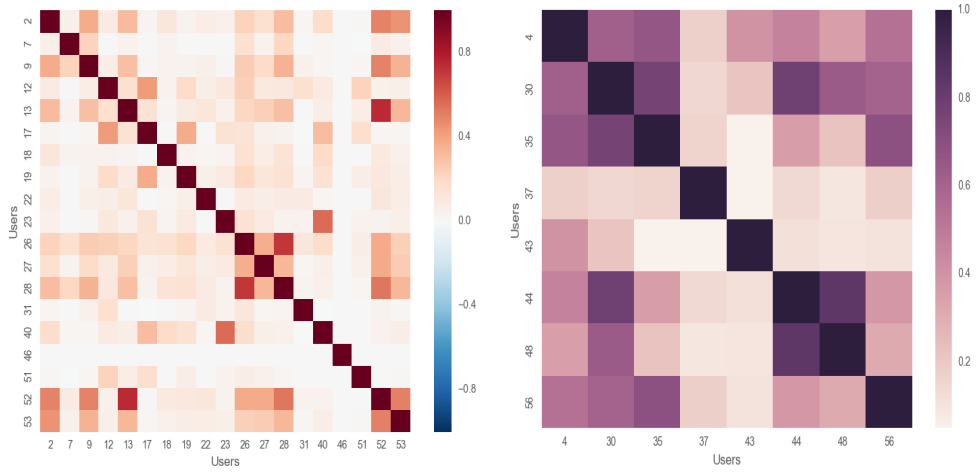
(a) MIT


(b) France


(c) Brazil

Figure 6.5. Sample interests for MIT traces (synthetic) and MACACO (web interests) for France and Brazil groups generated from their web browsing history using Interest-Learning Component of INDIGO.

users. Further $\mathbb{D}$ can also be restricted due to low interests similarities among mobile users or mobile user and data sources. Therefore, the maximum number of messages that can be stored in the network are $F = M * N * \mathbb{D}$. Every mobile user $u_j$ maintains a list $MList_{u_j}(t) = \{msg_i, \forall i \in [1, M]\}$ of all messages it receives up to time $t$ according to its mobility and interests. All mobile users $U$ can collect data directly from data sources or from mobile users using Contact & Data Gathering process (CDG) described as follows:




(a) France (b) Brazil

Figure 6.6. Interests similarities between volunteers of France and Brazil Groups calculated from their web interests profiles.

*Definition 1 (Contact & Data Gathering process (CDG): When any two mobile users or any mobile user or data source come in contact with each other they exchange their respective message lists if an only if they share similar interests (i.e utility value is 1).*

It is important to note that any two data sources *do not* exchange any data messages among them. Finally, I define multi-contact interest-based data dissemination process as follows:

*Definition 2 (Multi-contact Interest-based Data Dissemination): Any mobile user $u_j \in U$ gathers data message $msg_i$ from data source $d_i \in D$ or further disseminates its messages to other users using CDG process. The data dissemination process continues until all mobile users collect data source messages based on the overall data requirement $\mathbb{D}$. The mobility of mobile users, their interests similarities, multiple simultaneous contacts and data sources drives the Multi-contact Interest-based Data Dissemination.*

Algorithm 8 presents how every mobile user $u_j \in U$ gathers data at time $t$ using *CDG* process. $t^-$ and $t^+$ represent the time before and after $t$ respectively. I define data dissemination time as the time at which all users in the network fulfills data requirement $\mathbb{D}$. Consider a matrix $D^{ALL}(t)$ of size $N \times M$ that represents

**Algorithm 8** Contact & Data Gathering process (CDG)

**if** Any two users $u_j$ and $u_k$ come in contact at time $t$ with $MList_{u_j}(t^-)$ and $MList_{u_k}(t^-)$ **then**
  **if** $ut_{u_j u_k} = 1$ **then**
    Users $u_j$ and $u_k$ exchange all of their data messages
    $MList_{u_j}(t^+) = MList_{u_j}(t^-) \cup MList_{u_k}(t^-)$
    $MList_{u_k}(t^+) = MList_{u_j}(t^-) \cup MList_{u_k}(t^-)$
  **end if**
**end if**
**if** Any user $u_j$ with $MList_{u_j}(t^-)$ and any data source $d_i$ with message $msg_i$ come in contact at time $t$ **then**
  **if** $ut_{u_j d_i} = 1$ **then**
    Users $u_j$ collects data message $msg_i$ from $d_i$
    $MList_{u_j}(t^+) = MList_{u_j}(t^-) \cup msg_i$
  **end if**
**end if**

the list of all data messages collected up to time $t$ by $N$ mobile users.

$$D^{ALL}(t) = \begin{bmatrix} msg_{u_1 d_1} & msg_{u_1 d_2} & \dots & msg_{u_1 d_M} \\ msg_{u_2 d_1} & msg_{u_2 d_2} & \dots & msg_{u_2 d_M} \\ . & . & . & . \\ . & . & . & . \\ msg_{u_N d_1} & msg_{u_N d_2} & \dots & msg_{u_N d_M} \end{bmatrix}_{N \times M}$$

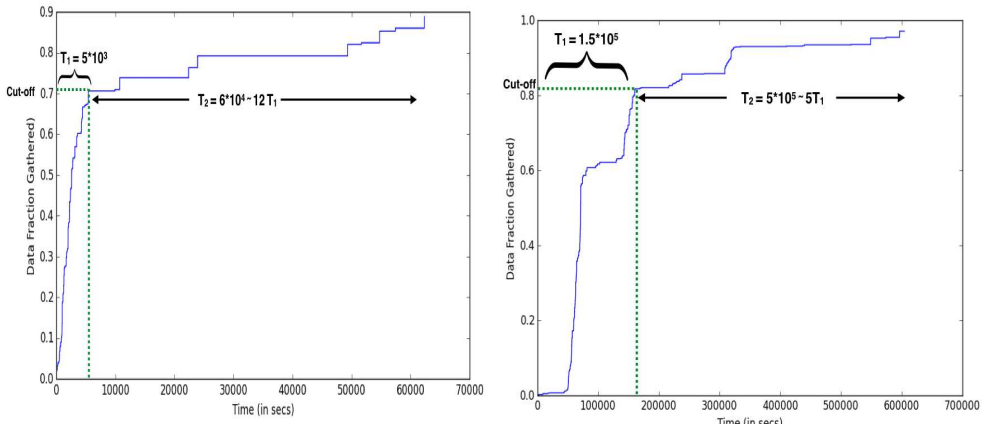
$$msg_{u_j d_i} = \begin{cases} 1 & if \quad msg_i \in MList_{u_j}(t) \\ 0 & \text{otherwise} \end{cases}$$

$$\forall\ j \in [1, N],\ \forall\ i \in [1, M]$$

The upper bound of data dissemination time $T_{dss^I}^{upper}$ is the maximum time slot at which data requirement will be fulfilled and $F$ elements of matrix $D^{ALL}(t)$ becomes 1.

### 6.4.2 Cut-off point approach for static and time-varying contact patterns

To verify the significance of Cut-off point $\alpha$ under interest-driven data dissemination strategy, I also analyze the fraction of data collected over time under




(a) INFOCOM Trace (b) MIT Trace

Figure 6.7. The fraction of data gathered with respect to time for all mobile users from all data sources for real-world traces from two diverse environments under interest-driven data dissemination strategy.

interest-driven data dissemination strategy from different real-world traces from diverse environments (conference Scott et al. [2006c] and university environment Eagle et al. [2009] during different time periods in Figure 6.7). From Figure 6.7 I observe the pattern similar to broadcast strategy: initially the rate of the data gathering is faster and after a certain data fraction, it exhibits a long tail cut-off. This happens because after a while the probability of getting new messages from neighboring users reduces due to extremely low contact with new users. In addition, the rate of data gathering depends on the inter-contact time among people that, after a power law period, exhibits a long tail cut-off Karagiannis et al. [2007]. Therefore, the time taken to gather remaining data increases significantly.

To determine $\alpha$, the `Markov Model` communicates with `Cut-off Estimator` by sending the fraction of data collected in current and previous step. For example in Figure 6.7, the estimated $\alpha$ for INFOCOM and MIT traces are 0.7 and 0.82 respectively. Once $\alpha$ is determined, the `Markov Model` models sets target state $S(F)$ according to user data requirement $\mathbb{D}$ and predicts $T_{dss^I}^{upper}$ using the *Fast Growing* and *Long Tail* phases of Cut-off based approach. The time spent in the *Fast Growing* and *Long Tail* phases are $T_{FGP^I}$ and $T_{LTP^I}$ respectively. $T_{FGP^I}$ represents the maximum time spent in Fast Growing phase (i.e. the maximum time required to reach directly to cut-off point state $S(\alpha * F)$ from $S(0)$). $T_{LTP^I}$ is the time spent in Long Tail Phase (i.e. maximum time spent in each state starting from $S(\alpha * F)$ to $S(F)$). Finally, the predicted upper bound of data dissemination

time $T_{dss^I}^{upper}$ under interest-driven data dissemination strategy can be defined as follows using $T_{FGP^I}$ and $T_{LTP^I}$:

$$T_{dss^I}^{upper} = T_{FGP^I} + T_{LTP^I} \tag{6.3}$$

Where,

$$T_{LTP^I} = \sum_{k=x}^{F} T_x \tag{6.4}$$

Long tail cut-off contributes higher weight to $T_{dss^I}^{upper}$ as compared to the time spend in gathering initial data messages and closely resembles the real-world data gathering process. To transit from one state to another, the `Markov Model` computes the transition probability $P_{S(x),S(x+k)}$ to reach $S(x+k)$ from $S(x)$ using contact probability matrix $P^C$ and utility matrix $\mathbb{UT}$. The transition probability for both static and time-varying cases is defined in Equation 6.5 and 6.6 respectively.

$$P_{S(x)S(x+h)} = \sum_{i \in S(x), j \in S(x+h)} p_{ij}^{c} \times ut_{ij} \tag{6.5}$$

$$P_{S(x)S(x+h)}^{k} = \sum_{i \in S(x), j \in S(x+h)} p_{ij}^{c_{day_k}} \times ut_{ij} \tag{6.6}$$

$$\text{Where } \forall\ p_{ij}^{c} \in P_S^C,\ \forall\ p_{ij}^{c_{day_k}} \in P_{day_k}^C,\ \forall\ P_{day_k}^C \in P_{TV}^C\ \forall\ u_{ij} \in \mathbb{UT}$$

Where $P_{S(x)S(x+h)}$ presents the transition probability to reach state $S(x)$ to $S(x+h)$ for static contact patterns case while $P_{S(x)S(x+h)}^{k}$ corresponds to the transition probability to reach state $S(x)$ to $S(x+h)$ for any $kth$ day i.e. $day_k, k \in [1,7]$ in time-varying contact patterns case under interest-driven data dissemination strategy. Once we reach the target state $S(F)$, the transition probability to remain in the same state will be 1. Since we also consider multiple simultaneous contacts among people, our Markov chain can directly jump to any state $S(x+n)$ from $S(x)$. This procedure helps to provide tighter prediction of $T_{dss^I}^{upper}$ by eliminating the impact of maximum time spent in each intermediate step.

Algorithm 9 presents our modified algorithm for interest-driven data dissemination to predict tighter upper bound for $T_{dss^I}^{upper}$ by computing $T_{FGP^B}$ and $T_{LTP^B}$ individually for both static and time-varying contact patterns using Algorithm 10 and Algorithm 11 respectively. Algorithm 10 and Algorithm 11 present the methodology to find maximum time ($maxTime$) spent in a state (i.e maximum time taken to reach $endState$ from $startState$ for several runs) under interest-based data dissemination by utilizing $P_S^C$ and $\mathbb{UT}$ for static contact patterns case and $P_{TV}^C$ for time-varying contact pattern case.

**Algorithm 9** Cut-off point based approach for prediction of $T_{dss^I}^{upper}$ for both static and time-varying contact patterns

**Require:** $P_S^C$, $P_{TV}^C$, $\mathbb{UT}$, $M > 0$,$N > 0$,$D^{ALL}(t) = 0_{N \times M}$, $NumRuns$,$F$,$MaxStepsDay$
**Ensure:** $T_{dss^I}^{upper}$
  $startState \leftarrow 0$
  $stopState \leftarrow F$
  $cutoffState \leftarrow 0$
  $T_{FGP^I} \leftarrow 0$, $T_{LTP^I} \leftarrow 0$, $T_{dss^I}^{upper} \leftarrow 0$, $\alpha \leftarrow 0$
  $T_{runs} = \emptyset$
  {estimate $\alpha$ from Cut-off Estimator Component}
  Set Cut-off point $\alpha$
  $cutoffState \leftarrow \alpha * F$ {calculate time for Fast Growing phase}
  **for** $i = 0$ **to** $NumRuns$ **do**
    **if** $STATIC\_CP\_CASE$ **then**
      $maxTime \leftarrow getMaxTimeStatic(startState, cutoffState, D^{ALL}(t), P^C, \mathbb{UT})$
      {For static contact patterns}
    **else**
      $maxTime \leftarrow getMaxTimeTV(startState, cutoffState, D^{ALL}(t), P_{TV}^C, MaxStepsDay, \mathbb{UT})$
      {For time-varying contact patterns}
    **end if**
    $T_{runs} \leftarrow T_{runs} \cup maxTime$
  **end for**
  $T_{FGP^I} \leftarrow max(T_{runs})$
  $T_{runs} = \emptyset$
  {Calculate time for Long Tail phase}
  $startState \leftarrow cutoffState$
  **while** $startState < stopState$ **do**
    $nextState \leftarrow startState + 1$
    **for** $i = 0$ **to** $NumRuns$ **do**
      $p \leftarrow startState$
      Randomly set $p$ elements of matrix $D^{ALL}(t)$ to 1
      **if** $STATIC\_CP\_CASE$ **then**
        $maxTime \leftarrow getMaxTimeStatic(startState, nextState, D^{ALL}(t), P^C, \mathbb{U})$
        {For static contact patterns}
      **else**
        $maxTime \leftarrow getMaxTimeTV(startState, nextState, D^{ALL}(t), P_{TV}^C, MaxStepsDay, \mathbb{U})$
        {For time-varying contact patterns}
      **end if**
      $T_{runs} \leftarrow T_{runs} \cup maxTime$
      **if** $maxTime = 0$ **then**
        $nextState \leftarrow nextState + 1$
      **else**
        $startState \leftarrow nextState$
        $nextState \leftarrow nextState + 1$
      **end if**
    **end for**
    $T_{LTP^I} \leftarrow T_{LTP^I} + max(T_{runs})$
  **end while**
  $T_{dss^I}^{upper} \leftarrow T_{FGP^I} + T_{LTP^I}$

**Algorithm 10** $getMaxTimeStatic(startState, endState, D^{ALL}(t), P_S^C, \mathbb{UT})$

**Require:** $startState$, $endState$, $D^{ALL}(t)$, $P_S^C$, $\mathbb{UT}$, $NumTrials$
**Ensure:** $maxTime$

**for** $trial = 1$ **to** $NumTrials$ **do**
  $T = \emptyset$
  $t = 0$
  **repeat**
    $x \leftarrow \mathscr{U}(0,1)$
    $t \leftarrow t + 1$
    {Assuming *CDG* process, user-user}
    **if** $x < 0.5$ **then**
      $r \leftarrow \mathscr{U}(0,1)$
      **for all** pairs of users $(i,j) : p_{i,j}^c \geq r$ AND $u_{ij} = 1$ **do**
        update $D^{ALL}(t)$ based on $CDG$ algorithm
      **end for**
    **end if**
    {Assuming *CDG* process, user-data source}
    **if** $x > 0.5$ **then**
      $r \leftarrow \mathscr{U}(0,1)$
      **for all** user and data source pair $(i,k) : p_{i,k}^c \geq r$ AND $u_{ik} = 1$ **do**
        update $D^{ALL}(t)$ based on $CDG$ algorithm
      **end for**
    **end if**
    $d \leftarrow count(D^{ALL}(t))$ {total messages stored in $D^{ALL}(t)$}
  **until** $d \leq endState$
  **if** $d = endState$ **then**
    $T \leftarrow T \cup t$
  **end if**
**end for**
$maxTime \leftarrow max(T)$
**return** $maxTime$

**Algorithm 11** $getMaxTimeTV(startState, endState, D^{ALL}(t), P_{TV}^{C}, MaxStepsDay, \mathbb{UT})$

**Require:** $startState$, $endState$, $D^{ALL}(t)$, $P_{TV}^{C}$, $\mathbb{UT}$, $NumTrials$,$dayNum$,$MaxStepsDay$
**Ensure:** $maxTime$

**for** $trial = 1$ **to** $NumTrials$ **do**
  $T = \emptyset$
  $t = 0$
  $tlocal = 0$
  $dayNum = 0$
  **repeat**
    $x \leftarrow \mathcal{U}(0,1)$
    $t \leftarrow t + 1$
    {Restart from the beginning contact probability if we reach $MaxStepsDay$}
    **if** $tlocal > MaxStepsDay \times \left|P_{TV}^{C}\right|$ **then**
      $tlocal \leftarrow 0$
    **end if**
    $tlocal \leftarrow tlocal + 1$
    {Get $P_{day_{dayNum}}^{C}$ for the given day using $tlocal$ and $MaxStepsDay$}
    $P_{day_{dayNum}}^{C} \leftarrow P_{TV}^{C}[dayNum]$
    {Assuming *CDG* process, user-user}
    **if** $x < 0.5$ **then**
      $r \leftarrow \mathcal{U}(0,1)$
      **for all** pairs of users $(i,j) : p_{i,j}^{c_{dayNum}} \geq r$ AND $u_{ik} = 1$ **do**
        update $D^{ALL}(t)$ based on $CDG$ algorithm
      **end for**
    **end if**
    {Assuming *CDG* process, user-data source}
    **if** $x > 0.5$ **then**
      $r \leftarrow \mathcal{U}(0,1)$
      **for all** user and data source pair $(i,k) : p_{i,k}^{c_{dayNum}} \geq r$ AND $u_{ik} = 1$ **do**
        update $D^{ALL}(t)$ based on $CDG$ algorithm
      **end for**
    **end if**
    $d \leftarrow count(D^{ALL}(t))$ {total messages stored in $D^{ALL}(t)$}
  **until** $d \leq endState$
  **if** $d = endState$ **then**
    $T \leftarrow T \cup t$
  **end if**
**end for**
$maxTime \leftarrow max(T)$
**return** $maxTime$

Algorithm 9 presents how the Cut-off point approach computes $T_{FGP^I}$ and $T_{LTP^I}$ individually to find tighter upper bound for $T_{dss^I}^{upper}$ for both static and time-varying contact patterns. Once the `Markov Model` determines $\alpha$ through `Cut-off Estimator`, it segregates Fast Growing and Long Tail phase. In Fast Growing phase, the model sets $startState$ as $S(0)$ and $cutoffState$ as $S(\alpha * F)$ and predicts ($T_{FGP^I}$) to directly reach $cutoffState$ for several runs (i.e. $NumRuns$ using $getMaxTime$ method described in Algorithm 10 and Algorithm 11). Afterward, it starts Long Tail phase by initializing $startState$ to $S(\alpha * F)$ and $nextstate$ to $S(\alpha * F + 1)$ and predicts maximum time spent in $nextstate$. If the model cannot reach $nextstate$ due to the non-existence of *CDG* process then, $nextstate$ will be marked as non-reaching and $maxTime$ to reach this state will be set to 0. In this case, the model will again compute $maxTime$ by setting $nextstate$ to $S(\alpha * F + 2)$ while keeping the same $startState$ as $S(\alpha * F)$. On the other side, if model could reach $S(\alpha * F + 1)$ (i.e $maxTime \neq 0$) then, it sets $startState$ as $S(\alpha * F + 1)$ and $nextstate$ as $S(\alpha * F + 2)$.

Likewise, the model repeats the same process until it reaches final state $S(F)$. Finally, the model computes $T_{LTP^I}$ as the sum of maximum time spent in each reachable state. As described in previous sections, the `Markov Model` predicts $T_{dss^I}^{upper}$ for different data requirements of the network starting from $\alpha$ to 1 unless $\mathbb{D}$ is provided as an input parameter.

## 6.5 Measured data dissemination time $T_{dss^I}^{meas}$ using real-world contact traces under interest-driven data dissemination strategy

To the best of my knowledge, in literature, there is no other work addressing the issue of heterogeneous contact patterns under interest-driven data dissemination strategy. Therefore, to compare the performance of upper bound of data dissemination time $T_{dss^I}^{upper}$ obtained from INDIGO framework, I simulate the real contact traces and measure the actual data dissemination time $T_{dss^I}^{meas}$ for each contact trace. I replayed all contact traces and measured $T_{dss^I}^{meas}$ for each trace under interest-driven data dissemination strategy. To have a comparable performance throughout the thesis, I consider the same segment of traces used in Chapter 5.

From the final traces, I randomly select $M$ users as mobile data sources and the remaining $N$ mobile users as those who are interested in gathering data messages from selected data sources according to different data requirements $\mathbb{D}$. I

update data matrix $D^{ALL}(t)$ based on the contact patterns observed in real contact traces and interest similarities among each pair of users. At each time step, I replay contact among a pair of users and impose interest-driven data dissemination strategy by comparing their interest similarity. If the interest similarity $sim_{ij}$ between a pair of users $i$ and $j$ is above $\beta$ threshold then the message will be exchanged. I stop the simulation when network data requirement gets fulfilled and record time as $T^{meas}_{dss^I}$. During these simulations of real-world contact traces, I also record the maximum possible fraction $DF_{max}$ of data that could be collected in the network by all nodes. To ensure comparable results, I also set maximum fraction of data collected $DF_{max}$ and measured data dissemination time $T^{meas}_{dss}$ through emulation of real traces as ground truth for the `Markov Model` of INDIGO. Please note that the $T^{meas}_{dss^I}$ will consist all real aspects of contact patterns i.e. time-varying contact patterns, multiple simultaneous contacts, and interests-driven data exchange. Therefore, I create the ground truth of data dissemination time $T^{meas}_{dss^I}$ for all traces and utilize it as a benchmark for both static and time-varying data dissemination prediction.

For all contact traces, in general, the maximum fraction of data collected is lesser under interest-driven data dissemination strategy as compared to the broadcast strategy. This observation shows that the interest-driven data dissemination strategy restricts spread of information in most of the weeks (max collected data fraction was less than 100%). The MACACO contact traces along with real interests also exhibit a similar trend. I observe that for France-$W^H$ and France-$W^L$ (the difference in the fraction of data collected is only slightly higher (5%)) even though France-$W^H$ has much higher contact probabilities among people as compared to France-$W^L$. Brazil group also produces similar effects as data fraction collected in Brazil-$W^H$ is only 10% higher as compared to Brazil-$W^L$. The $T^{meas}_{dss^I}$ results for both groups show the importance of interest-driven data dissemination strategy on data dissemination time. Table 6.2 presents simulation settings and $T^{meas}_{dss^I}$ for maximum data fraction $DF_{max}$ collected in all weeks.

## 6.6 Results and discussion

In this section, I validate the upper bound of data dissemination time $T^{upper}_{dss^I}$ obtained against the ground truth data dissemination time, $T^{meas}_{dss^I}$ under interest-driven data dissemination strategy from all real-world contact traces. For each selected contact trace segment of contact traces (same as used in Chapter 5), INDIGO predicts $T^{upper}_{dss^I}$ for all data requirements starting from $\alpha$ to $DF_{max}$ with 5% step size for both contact patterns using Cut-off based approach. For all con-

Table 6.2. Simulation settings and $T_{dss^I}^{meas}$ for maximum data fraction for all contact traces under interest-driven data dissemination strategy.

| **Trace** | # **Mobile Users** | #**Data Sources** | $DF_{max}$ | $T_{dss^I}^{meas}$ (in secs) |
|---|---|---|---|---|
| INFOCOM ($2^{nd}$ day) | 30 | 9 | 0.89 | 62311 |
| PERCOM ($4^{th}$ day) | 32 | 9 | 0.99 | 32420 |
| ROLLERNET | 50 | 12 | 1 | 1590 |
| MIT-W1 (from $10^{th}$ month) | 51 | 15 | 0.894 | 599428 |
| MIT-W2 (from $10^{th}$ month) | 57 | 15 | 0.826 | 584668 |
| MIT-W3 (from $10^{th}$ month) | 61 | 15 | 0.97 | 596697 |
| MIT-W4 (from $10^{th}$ month) | 64 | 15 | 0.96 | 603377 |
| MACACO–France–$W^H$ | 15 | 4 | 0.5714 | 308020 |
| MACACO–France–$W^L$ | 15 | 4 | 0.5238 | 542396 |
| MACACO–Brazil–$W^H$ | 6 | 4 | 0.50 | 361425 |
| MACACO–Brazil–$W^L$ | 6 | 4 | 0.40 | 215638 |

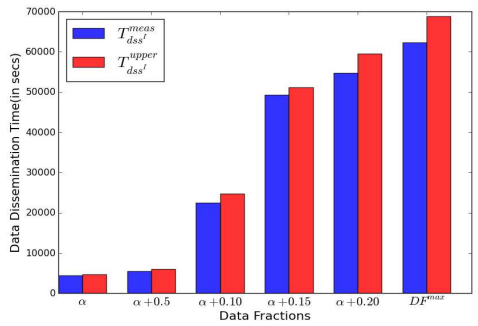
tact trace segments, the contact probabilities were predicted through the *Contact Probability Prediction Module* and the Cut-off point $\alpha$ was estimated with the help of `Markov Model` and `Cut-off Estimator` of *Interest-Driven* sub-module. Further, the interest threshold value $\beta$ is also set by the `Data Processor` for all contact traces using $\mathbb{E}(I^S)$. Based on $\beta$, the `Data Processor` prepares $\mathbb{UT}$ and provide it as an input to `Markov Model` of INDIGO to model interest-driven data dissemination strategy. The values for $\beta$ for INFOCOM, PERCOM, ROLLERNET, and MIT is set as 0.5, 0.6, 0.6, (W1, W2–0.5, W3, W4–0.4) respectively. Further, for France and Brazil group of MACACO traces the $\beta$ is set as 0.3 and 0.5 respectively. From $I^S$ matrix, we observe that volunteers in France share fewer interests as compared to volunteers in Brazil. The difference in interests is primarily because of the diversity of volunteers in terms of different departments and age groups. Each experiment is repeated for several runs for statistical convergence. Now I will present the results for $T_{dss^I}^{upper}$ for both static and time-varying contact probabilities for interest-driven data dissemination strategy.

### 6.6.1 Static contact probabilities case

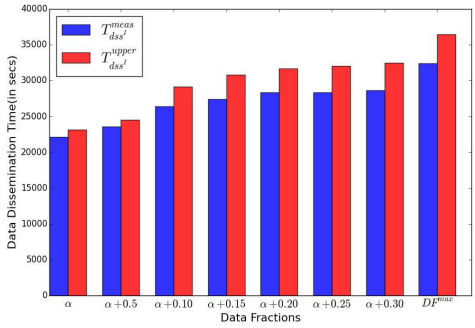
Figure 6.8 and Figure 6.9 present how well INDIGO is able to predict $T_{dss^I}^{upper}$ against $T_{dss^I}^{meas}$ for different contact traces under interest-driven data dissemination and different data requirements.

From these Figures, we observe that the Cut-off approach once again provides tighter prediction of $T_{dss^I}^{upper}$ for all contact traces under interest-driven data dissemination strategy. In addition to this, we also find the applicability of INDIGO for different data requirements starting from $\alpha$ to $DF_{max}$. For conference environment, we find the error between 4-13% and find comparable results for both INFOCOM and PERCOM trace and shows the significance of interest-driven data dissemination strategy. Even though the contact rate in INFOCOM trace among people is slightly higher than PERCOM still, due to high interest similarities among people, the likelihood to exchange more information increases in PERCOM whenever people meet with each other. For ROLLERNET trace of crowded urban environment, we also observe that the Cut-off based approach provides the tighter upper bound of data dissemination time $T_{dss^B}^{upper}$ between 11-13% error. As compared to conference environment, the higher prediction error in ROLLERNET is due to the overestimation of time spent in Long Tail phase while in reality people were more quickly disseminating information among each other as they were accumulated for a city tour.

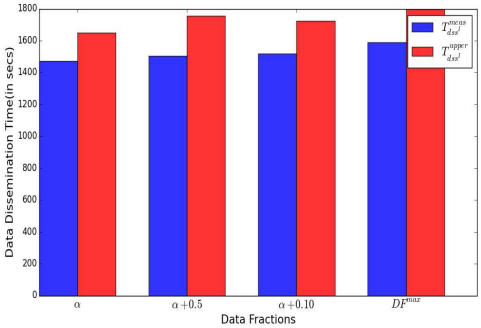
Figure 6.8(d), 6.8(e),6.8(f),6.8(g) depict the results obtained from MIT traces for all 4 weeks (MIT-W1 to MIT-W4) and once again shows the applicability of



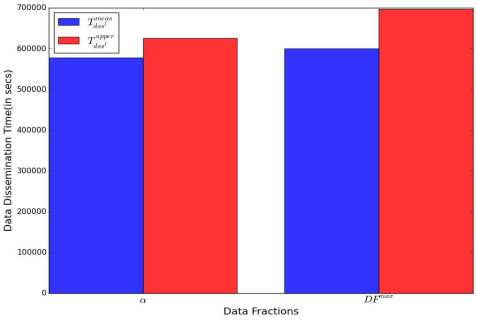

(a) INFOCOM–$\alpha = 0.65$



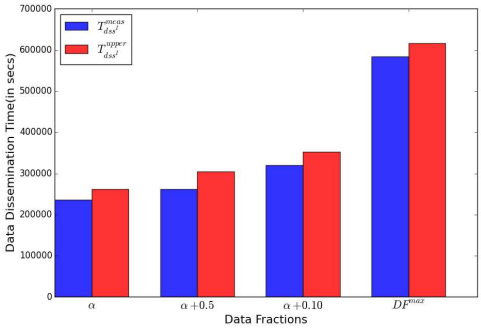

(b) PERCOM–$\alpha = 0.65$



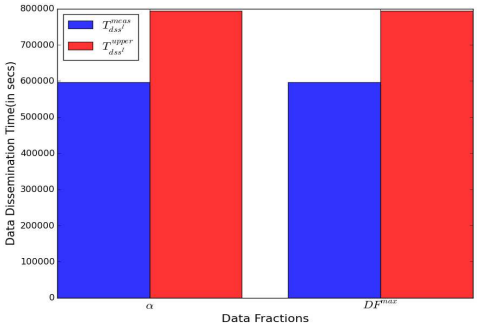

(c) ROLLERNET–$\alpha = 0.89$



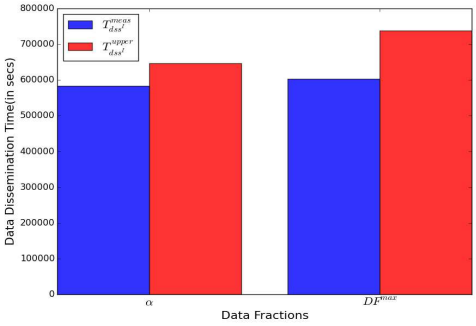

(d) MIT-W1–$\alpha = 0.87$




(e) MIT-W2–$\alpha = 0.68$

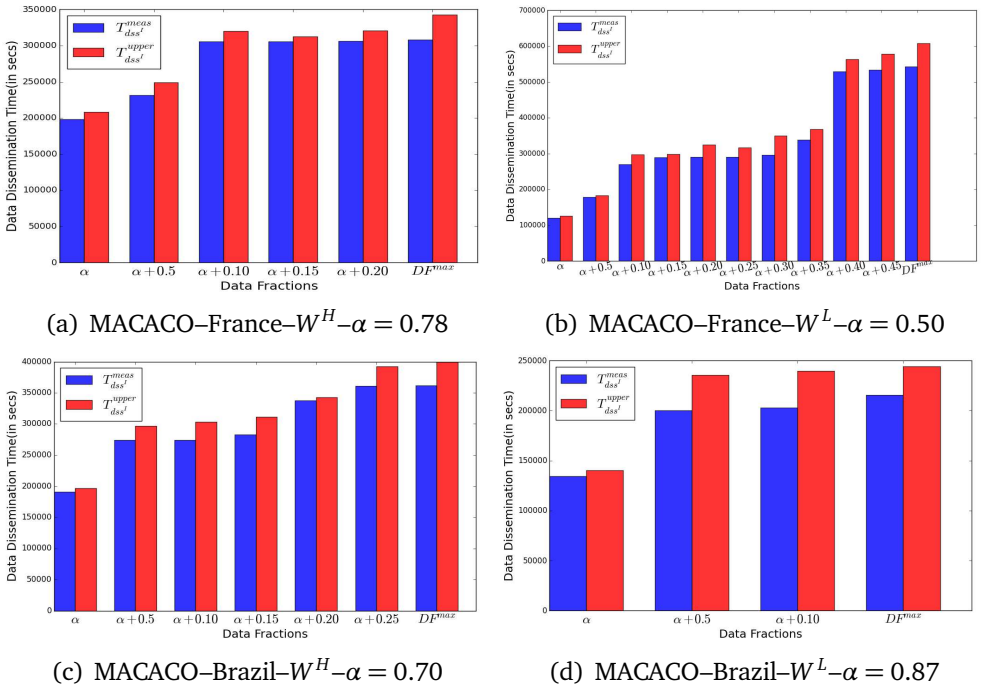



(f) MIT-W3–$\alpha = 0.96$




(g) MIT-W4–$\alpha = 0.943$

Figure 6.8. Comparison of real data dissemination time $T^{meas}_{dss^l}$ against the upper bounds of data dissemination time $T^{upper}_{dss^l}$ predicted using Cut-off approaches for INFOCOM, PERCOM, ROLLERNET, and MIT trace under interest-driven dissemination strategy and static contact probabilities.




(a) MACACO–France–$W^H$–$\alpha = 0.78$ (b) MACACO–France–$W^L$–$\alpha = 0.50$

(c) MACACO–Brazil–$W^H$–$\alpha = 0.70$ (d) MACACO–Brazil–$W^L$–$\alpha = 0.87$

Figure 6.9. Comparison of real data dissemination time $T^{meas}_{dss^I}$ against the upper bounds of data dissemination time $T^{upper}_{dss^I}$ predicted using Cut-off approaches for MACACO trace under interest-driven dissemination strategy and static contact probabilities.

INDIGO for long duration contact traces by providing tighter upper bound of $T^{upper}_{dss^I}$ under interest-driven data dissemination strategy. We also observe that for the first 2 weeks (MIT-W1 to MIT-W2), the upper bound $T^{upper}_{dss^I}$ is almost similar because the contact pattern and interests similarities of people are relatively identical (prediction error is between 5-16%). However, for the (MIT-W3 to MIT-W3) week, due to fewer contacts among people (semester break at MIT) and less interest-similarities, INDIGO obtains higher error in $T^{meas}_{dss^I}$ prediction (error lies between 11-33%). Similar to broadcast strategy, I also notice that $T^{upper}_{dss^I}$ is comparatively less tight in MIT than INFOCOM and PERCOM and ROLLERNET traces.

Finally, for MACACO traces all weeks of both groups, INDIGO predicts $T^{upper}_{dss^I}$ within 5-18% error for all data requirements starting from $\alpha$ to $DF_{max}$. For MACACO–France–$W^H$, it predicts $T^{upper}_{dss^I}$ within 2-11% error while for MACACO–France-$W^L$, the prediction of upper bounds gets looser with an error of 2-18%. This happens due to lower contact probabilities predicted from INDIGO and its

impact on Long Tail phase of Cut-off point based approach. We also observe similar trend in Brazil group with 2-15% and 13-18% error in MACACO–Brazil-$W^H$ and MACACO–Brazil-$W^L$ weeks respectively. The impact of interest similarities are also evident in MACACO trace due to comparable errors obtained in France and Brazil traces even though Brazil traces exhibits much lower contact probabilities. In the case of France traces, INDIGO realizes high contacts among people but do not exchange message due to lower interest similarities thus results in overall increases of $T_{dss^I}^{upper}$. However, for Brazil traces it exhibits lower contacts but a higher number of messages are exchanged due to high interest similarities among people. For all traces, our results show highest and lowest error at $DF_{max}$ and $\alpha$ data requirements respectively. This result once again signifies the impact of Long Tail phase on data dissemination time because, in the case of $\alpha$ data requirement, INDIGO only utilizes Fast Growing Phase while for others, it uses both Fast Growing and Long tail phase.

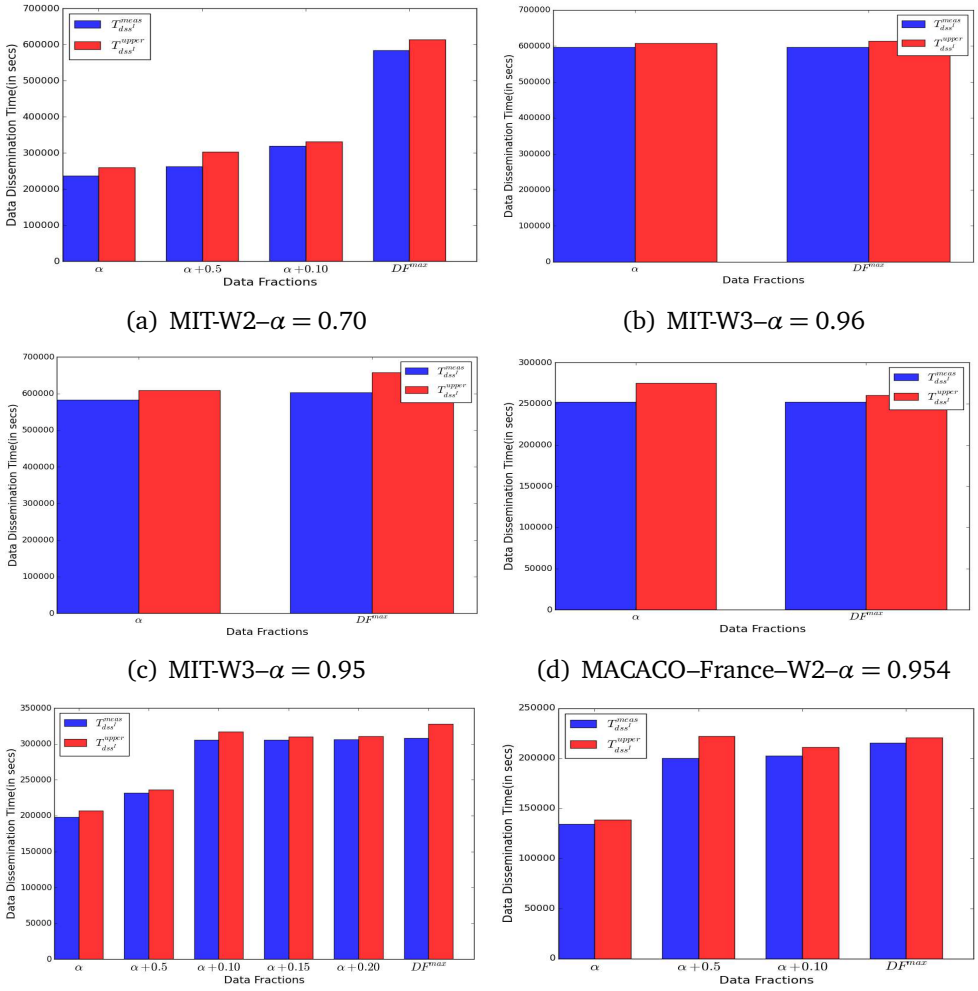
From the above discussion, I conclude that *Cut-off point* approach outperforms for the other strategy i.e. interest-driven data dissemination strategy for all traces under static contact patterns and different data requirements. The results also show the impact of interest similarities on the perdition of $T_{dss^I}^{upper}$ while showing the comparable performances for traces exhibiting lower contact probabilities.

### 6.6.2 Time-varying contact probabilities case

In this Section, I present the results of $T_{dss^I}^{upper}$ under time-varying contact patterns and interest-driven strategy for MIT and MACACO traces. As explained in previous Chapters, for time-varying contact probabilities, I only contact traces collected for a longer duration of time and exhibits certain regularities in their contact patterns.

Figure 6.10 presents the applicability of INDIGO in predicting $T_{dss^I}^{upper}$ against $T_{dss^B}^{meas}$ for the different weeks of MIT and MACACO traces for different data requirements. For MIT trace, INDIGO trains the contact probability prediction model with MIT-W1 contact probabilities followed by the pair-wise contact probabilities prediction for subsequent of the weeks. Similarly, for MACACO trace, the *Contact Probability Prediction Module* trains the model for the 1-week trace of both France and Brazil and predict the contact probabilities for future weeks. Once the contact probabilities for other weeks of MIT and MACACO traces are predicted then INDIGO predicts $T_{dss^I}^{upper}$ for all these weeks. The value of $\beta$ are same for all contact traces as in static contact patterns case.

From Figure 6.10, we can observe that time-varying contact probabilities fur-

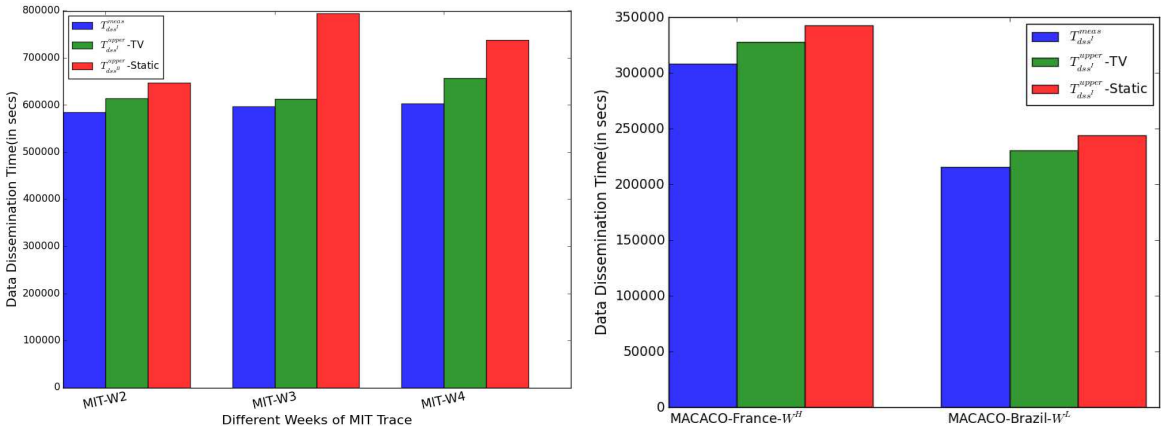



(a) MIT-W2–$\alpha = 0.70$ (b) MIT-W3–$\alpha = 0.96$

(c) MIT-W3–$\alpha = 0.95$ (d) MACACO–France–W2–$\alpha = 0.954$

(e) MACACO–France–W3 ($W^H$)–$\alpha = 0.82$ (f) MACACO–Brazil–W3 ($W^L$)–$\alpha = 0.89$

Figure 6.10. Comparison of real data dissemination time $T^{meas}_{dss^l}$ against the upper bounds of data dissemination time $T^{upper}_{dss^l}$ predicted using Cut-off approaches for MIT and MACACO traces under interest-driven dissemination strategy and time-varying contact probabilities.




(a) MIT (b) MACACO

Figure 6.11. Comparison of $T_{dss^I}^{meas}$ against $T_{dss^I}^{upper}$ predicted using for different weeks and maximum data fraction of MIT and MACACO traces under interest-driven dissemination strategy and both static and time-varying contact patterns.

ther tightens the upper bound of data dissemination time. For MIT traces, the error lies between 5-15% while for MACACO trace it is between 3-15%. We also observe the trend similar to broadcast strategy where the error prediction of $T_{dss^I}^{upper}$ decreases for subsequent weeks due to more accurate prediction of contact probabilities over time. The results also show increase in $\alpha$ for time-varying contact patterns case thus decrease in time spent in Long Tail phase and prediction error of $T_{dss^I}^{upper}$. The change in $\alpha$ is dynamically estimated by the `Cut-off Estimator`.

Finally, to show the comparative results of $T_{dss^I}^{upper}$ from both static and time-varying contact patterns, I altogether present the prediction of $T_{dss^I}^{upper}$ against $T_{dss^I}^{meas}$ for both contact patterns in Figure 6.11 for the maximum fraction of data $DF_{max}$. The Figure clearly shows that time-varying contact patterns exhibit more realistic contact patterns of people thus provides much tighter upper bound of $T_{dss^I}^{upper}$ as compared to the static contact patterns against the ground truth $T_{dss^I}^{meas}$.

## 6.7 Conclusions

In this Chapter, I presented the prediction of the upper bound of data dissemination time using the *Interest-Driven Sub-Module* of INDIGO framework under multi-source, multi-contact interest-driven data dissemination for both static and

time-varying contact patterns for Type III case. I proposed a method to learn the real interests of people from their browsing history and also a way to create artificial interests using the *Interest-Leaning Component*. I also presented the modified Cut-off based approach in INDIGO that predicts upper bound of data dissemination time by considering both learned real interests of people from web browsing history of their Smartphones along with their heterogeneous contact patterns.

The prediction results obtained for the upper bound of data dissemination time showed the applicability of INDIGO for interest-driven data dissemination strategy for both static and time-varying contact patterns by providing the tighter bounds. I validated INDIGO through our real-world traces collected and achieved tight upper bounds of data dissemination for both types of interests (synthetic as well as real) and contact patterns. More specifically, I observed that employing time-varying contact patterns improves the tighter bounds of data dissemination time as they reflect more realistic contact patterns of people. I also summarize the key observation from this Chapter as follows:

- To the best of my knowledge, INDIGO is the first work that predicts tighter upper bound of data dissemination time by learning real interests of people from web browsing history of their Smartphones along with heterogeneous mobility aspects. Further to the best of knowledge, the MACACO traces are the first trace set that captures interests of people along with their real mobility.

- Similarly to inter-contact time I identify a power-law behavior with a long tail cut-off that, to the best of my knowledge, has not been identified before. Such discovery is very important as it has a strong impact on the data dissemination time.

- I exploit the previous finding by providing tighter upper bound of data dissemination time by using the Cut-off point based approach for both static and time-varying contact patterns. Prediction from time-varying contact patterns improves with time due to more training data.

- Interest-driven data dissemination strategies are effective to restrict the spread of dissemination as opposed to broadcast strategy because information is only disseminated for people with sharing similar interests.

- The use of time-varying contact patterns, instead of static ones, leads to an increase in $\alpha$ that results in the reduction of the prediction error of $T_{dss^I}^{upper}$.

The work done in this Chapter are under submission at IEEE WoWMoM 2017 for modeling of interest-driven data dissemination and the interest-learning part

done in Telefónica research is under ACM Transaction of Web (TWEB) 2017. Till now, I have focused on the prediction of the upper bound of data dissemination time using INDIGO for both Type II and Type III cases. In next Chapter, I will focus how to find the best relays in the network utilizing the *BROP* model of INDIGO for both types and how does it impacts the data dissemination time.

# Chapter 7

# Estimation of Best Relays Using BROP Model

## 7.1 Introduction

In previous Chapters, I presented the modeling and prediction of data dissemination process for the tighter upper bound of data dissemination time under different dissemination strategy and contact patterns. In this Chapter, I will focus on another dimension of data dissemination process i.e. finding the best relays in the network to speed up the information diffusion in the network for Type II and III cases of physical networks. In this way, INDIGO can help local businesses to target those people who can spread their local services and advertisements much quicker while covering more people.

Searching for best spreaders in complex networks is an issue of great significance for applications across various domains, ranging from the epidemic control Anderson et al. [1992] Heesterbeek [2000] Pastor-Satorras and Vespignani [2001], viral marketing Watts et al. [2007] Leskovec et al. [2007] and social movement to idea propagation Diani and McAdam [2003] Lü et al. [2011] Myers et al. [2012b] Zhang et al. [2016]. To find the super spreaders in such complex networks, these works focuses on the network properties using centrality measures like degree centrality (or just the degree of a node, i.e. the number of its links), the eigenvector centrality Bonacich [1987], the betweenness centrality Freeman [1977] etc. Recently, the another centrality measure based on the notion of K-cores is applied in many real networks Dorogovtsev et al. [2006] Carmi et al. [2007] Garas et al. [2010] and shown to be effective in understanding network structure and finding influential nodes in the network Batagelj and Zaveršnik [2011].

One of the major limitation of the above-described centrality measures, including the K-core decomposition method, is their design to work on unweighted graphs. However, in practice, real networks have weights that describe important and well-defined properties between the graph nodes. To handle such complex networks, the authors in Garas et al. [2012] proposed a weighted K-Shell decomposition algorithm that takes into account both the degree centrality and weight measures to find best relays in the network. In the case of INDIGO, the contact strength (physical proximity) and interest similarity (social proximity) among people can also be represented as a weighted graph and I need to address the problem of finding best relays from these weighted graphs. Therefore, to find the set of *Best Relays*, INDIGO also utilizes the weighted K-shell decomposition algorithm for Type II and III cases.

The chapter is structured as follows. In Section 7.2, I will present different parts and working of *BROP Component* to find the best relays in the network under broadcast and interest-driven data dissemination strategy. Further in Section 7.3, I will describe the weighted k-shell decomposition algorithm and how do I use it for broadcast and interest-driven strategy along with their impact on data dissemination time. Afterward, in Section 7.4, I will discuss the results obtained from *BROP Component* for both broadcast and interest-driven data dissemination for different traces. Finally, with Section 7.5, I conclude this Chapter.

## 7.2 Overview of BROP Component

Figure 7.1 and 7.2 presents the detailed view of *BROP Component* which is responsible for finding *Best Relays* in the network under both broadcast and interest-driven data dissemination strategy. For broadcast strategy, the *BROP Component* inputs pair-wise contact probabilities predicted through the *Contact Probability Prediction Module* and further finds the *Best Relays* in the network by detecting the Core and Non-Core nodes using K-Shell sub-component. The Core nodes represent the super spreaders of the network while Non-Core nodes represent those nodes who are not central in the network (more details in next section). All input given to the K-Shell sub-component is provided by the Data Processor. In the case of interest-driven data dissemination strategy, (see Figure 7.2) in addition to contact probabilities, the interest-similarities among people learned from the *Interest Learning Component* are also given as input to the Data Processor. Once the Core and Non-Core nodes of the network are identified then, based on the *data requirement*, *# of data sources* and *interest similarity threshold*, the *BROP Component* selects different sets of initial data sources and

finds the optimum set of *Best Relays* that minimizes the upper bound of data dissemination time.

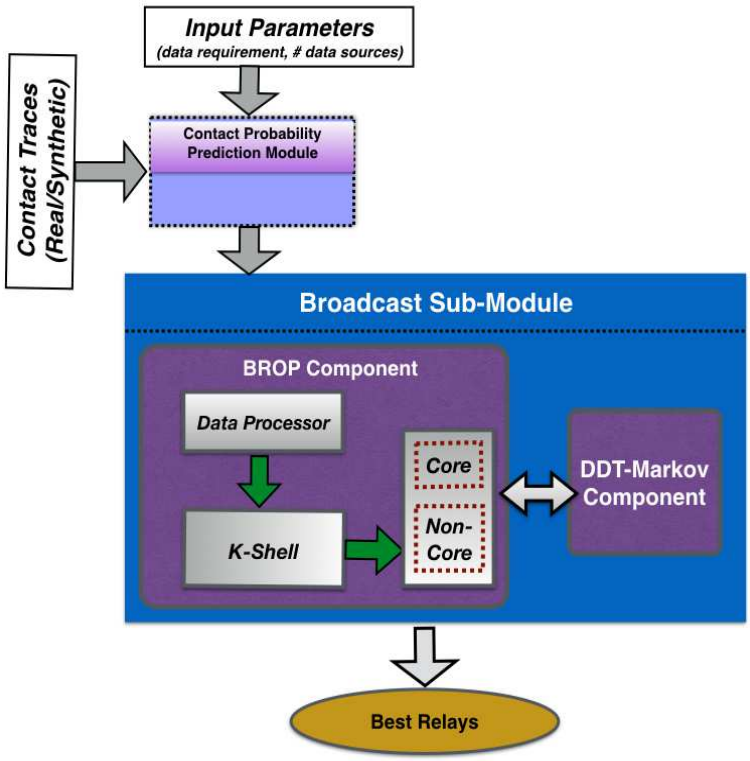



Figure 7.1. BROP Component of INDIGO required to find Best Relays under broadcast data dissemination strategy for Type II case.

### 7.2.1 Broadcast data dissemination strategy

The working of *BROP Component* for broadcast data dissemination strategy is presented in Figure 7.3. In this case, the pair-wise contact probabilities and other parameters are given as an input to the Data Processor that builds the network by finding the unique nodes and create a weighted edges $w_{ij}^{B}$ between each node pair $i, j$ using their contact probability $p_{ij}$. Once the network is built, it is passed on to the K-Shell sub-component that employs the weighted K-Shell decomposition algorithm on a network to finds the *Core* and *Non-Core* nodes by considering both node degree and associated weights with other nodes in the network. Further, to find the best-suited set of *Best Relays*, the *BROP Component* communicates with *DDT-Markov Component* by sending a set of *Core* nodes (selected based on the #

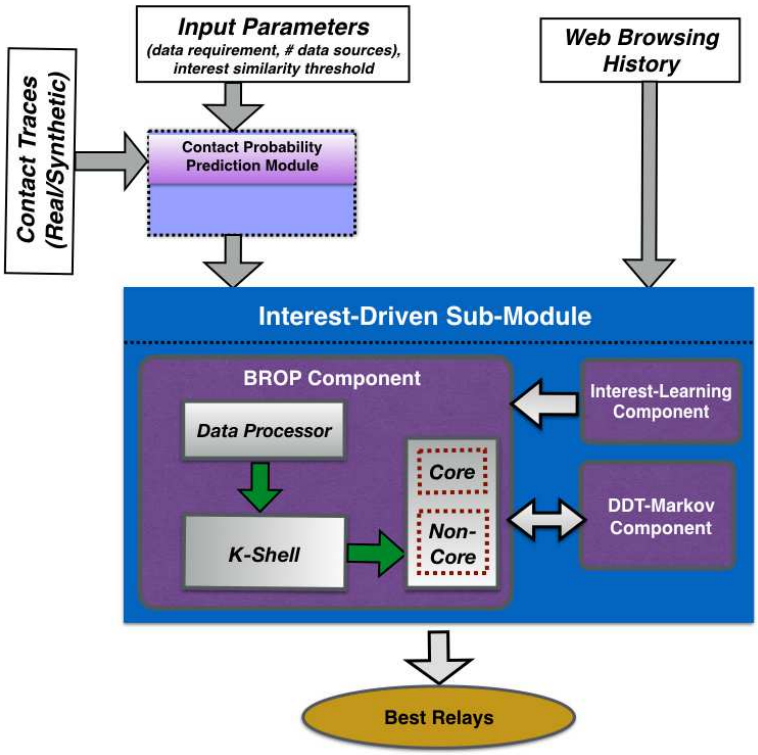



Figure 7.2. BROP Component of INDIGO required to find Best Relays under interest-driven data dissemination strategy for Type III case.

*of data sources*) and finds the *Best Relay* set that minimizes the data dissemination time for a certain *data requirement*.

### 7.2.2 Interest-driven data dissemination strategy

As presented in Figure 7.4, the *BROP Component* under interest-driven strategy also takes pair-wise contact probabilities as an input along with pair-wise interest similarities among people obtained through *Interest Learning Component*. To consider the impact of interest-driven data dissemination strategy, the *BROP Component* eliminates the edges between those pair of nodes $i, j$ whose utility value $u_{ij}$ is 0 while the building of network (utility value description is for interest-driven data dissemination is discussed in detail in Chapter 6). Afterward, it applies weighted K-Shell decomposition on the network through K-Shell sub-component and finds the set of *Best Relays* using *DDT-Markov Component* of INDIGO's `Interest-Driven Sub-Module`.

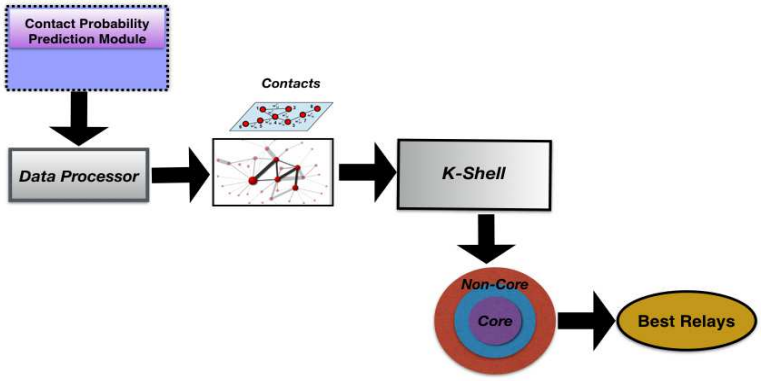



Figure 7.3. Working of BROP Component to find best relays using K-Shell under broadcast data dissemination strategy.

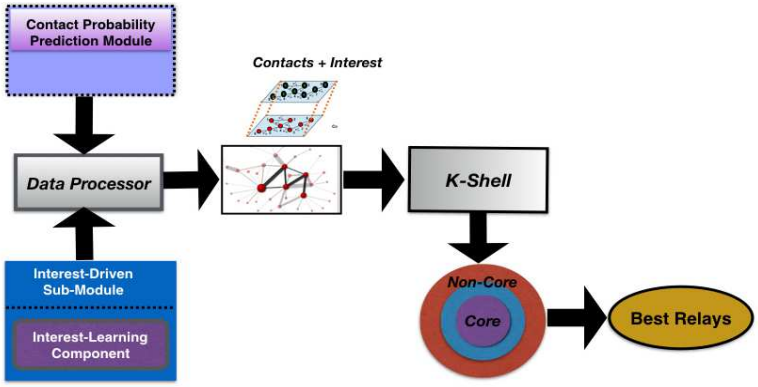



Figure 7.4. Working of BROP Component to find best relays using K-Shell under interest-driven data dissemination strategy.

## 7.3 Weighted K-Shell decomposition algorithm

In this Section, I will give the overview of weighted K-Shell decomposition algorithm used by the *BROP Component* of INDIGO framework. Weighted K-Shell decomposition algorithm is an extension of unweighted K-Shell decomposition algorithm where it considers both degree of a node and the weights of its links. The traditional K-Shell decomposition method partitions a network into sub-structures that are directly linked to centrality Alvarez-Hamelin et al. [2005] Batagelj and Zaveršnik [2011]. This method assigns an integer index, $k_s$, to each node that is representative of the location of the node in the network, according to its node degree. Nodes with low values of $k_s$ are located to the periphery of the network while nodes with a high value of $k_s$ will reside in the

center of the network. This way, the network is described by a layered structure (similar to the structure of an onion), revealing the full hierarchy of its nodes. The innermost nodes belong to the structure called "Core" of the network, while the remaining nodes are placed into more external layers (K-Shells). Figure 7.5 presents the how a network is divided into this K-Shell structure using the K-Shell decomposition algorithm. In the beginning, the algorithm first recursively removes all nodes with degree $K = 1$ from the network and assigns the integer value $k_s = 1$ to them. This procedure is repeated iteratively until there are only nodes with degree $K \geq 2$ left in the network. Subsequently, it removes all nodes with degree $K = 2$ and assigns to them the integer value $k_s = 2$. Again, this procedure is repeated iteratively until there are only with nodes with degree$K \geq 3$ left in the network, and so on. This routine is applied until all nodes of the network have been assigned to one of the K-Shells.

The above described original K-Shell decomposition does not consider the weights of the links. To find the *Best Relays* in a network, the *BROP Component* of INDIGO needs to take into account the physical (contact) and social (interest similarity) weights among people. Therefore to address this problem, I also applied weighted K-Shell decomposition algorithm proposed by Garas et al. [2012] in *BROP Component*. The weighted K-Shell decomposition algorithm applies the same pruning routine described earlier, but it is based on an alternative measure for the node degree. This measure considers both the degree of a node and the weights of its links. Considering these two measures, the algorithm assigns a weighted degree, $k'_i$ for a node $i$ is defined as:

$$k'_i = \left[ k_i^{\theta} \left( \sum_{j}^{k_i} w_{ij} \right)^{\gamma} \right]^{\frac{1}{\theta+\gamma}} \tag{7.1}$$

where $k_i$ is the degree of node $i$, and $\sum_j^{k_i} w_{ij}$ is the sum over all its link weights. The value of $\theta$ and $\gamma$ determines the importance of degree and weights. The authors of weighted K-Shell decomposition algorithm have considered both $\theta$ and $\gamma$ values as 1 to give equal importance to degree and weights. Similar to them, I will also set $\theta$ and $\gamma$ values as 1 for the *BROP Component*. Finally, the Core nodes obtained using *BROP Component* are considered as Best Relays of the network to enable faster information dissemination in the network. The *BROP Component* marks Core and all Non-Core nodes as 0 and 1 respectively.

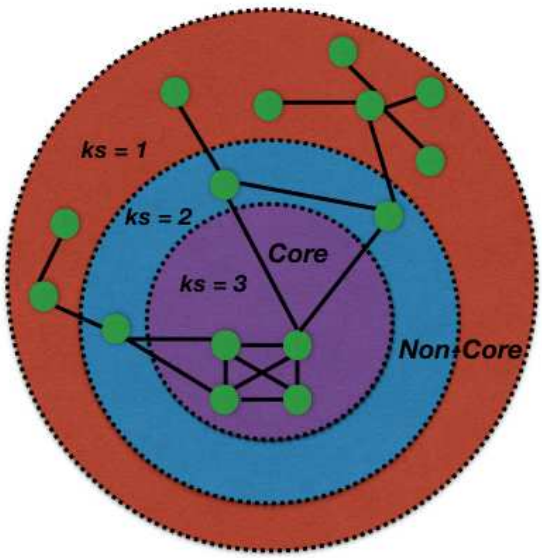


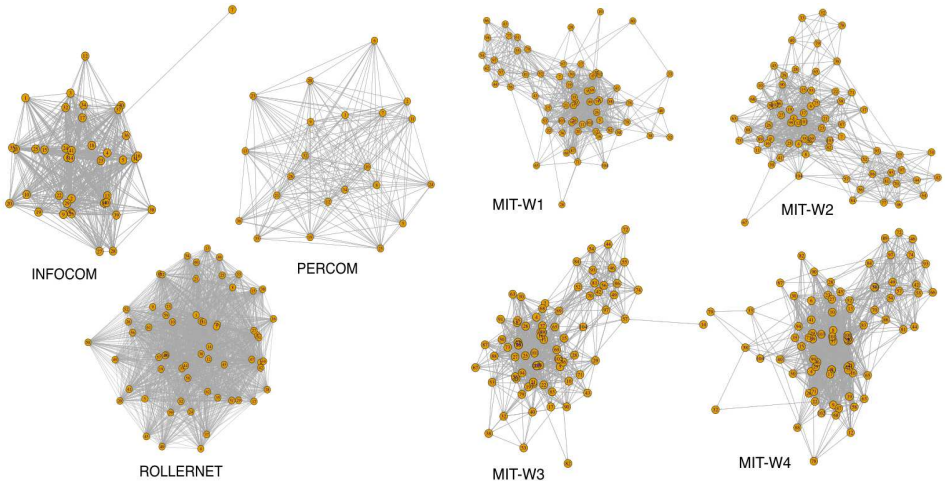

Figure 7.5. Illustration of the layered structure of a network obtained using the k-shell decomposition method. The nodes between the two outer rings include nodes of shell 1 ($ks = 1$), while the nodes between the two inner rings compose shell 2 ($ks = 2$). The nodes within the central ring called as Core ($ks = 3$) while nodes of other outer and middle rings are called as Non-Core nodes ($ks = 1, 2$).

### 7.3.1 Impact of Core and Non-Core nodes on data dissemination time

In this Section, I will present the impact of Core and Non-Core nodes on data dissemination time for different traces. Figure 7.6 and Figure 7.7 present the structure of network[1] and sample plots of the mean of the data dissemination times obtained from the different set of Core and Non-Core nodes using *BROP Component*.

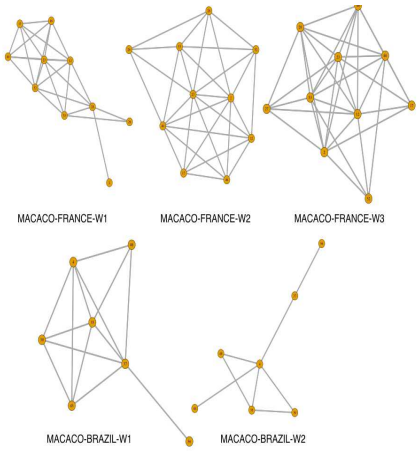
From Figure 7.6, we observe that both conference environment traces (INFOCOM and PERCOM) have a connected community due to the gathering of people for the conference event. Further, for the urban environment ROLLERNET trace, we observe a complete tight knitted community with high intensity of contacts. This happens because people were on a city tour and walking altogether. For all weeks of MIT trace, we clearly see two groups: the first one is the group of researchers in a department and the another one is a set of visiting researchers. For MACACO trace, we also see a community formed from the

[1]For the sake of convenience, I gave some pseudo name to the nodes in the network.



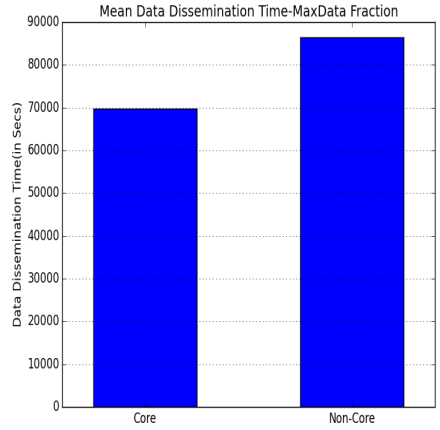

(a) INFOCOM, PERCOM and ROLLERNET

(b) All weeks of MIT

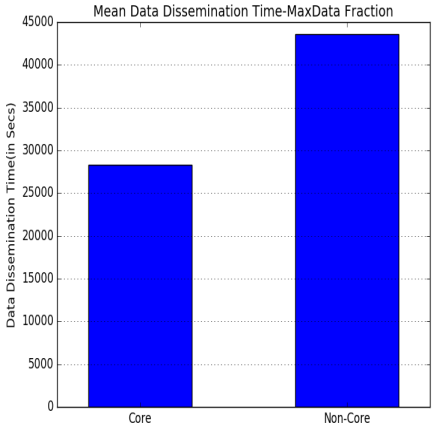



(c) All weeks of MACACO for both France and Brazil

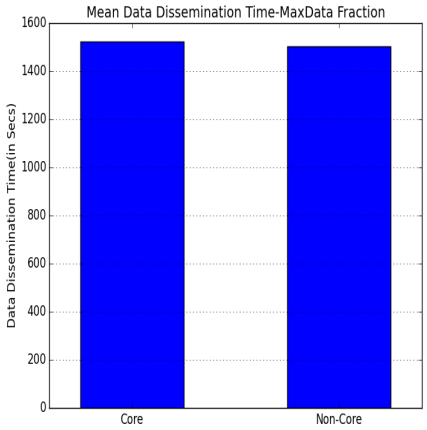
Figure 7.6. Contact graph of the network for each trace.



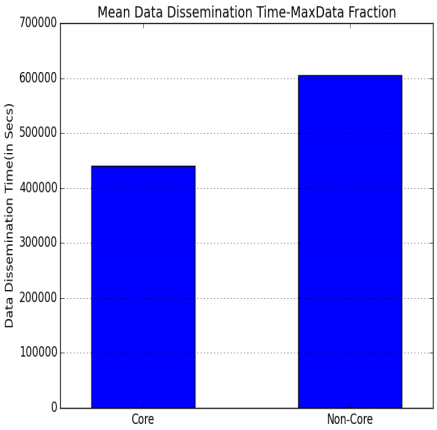

(a) INFOCOM



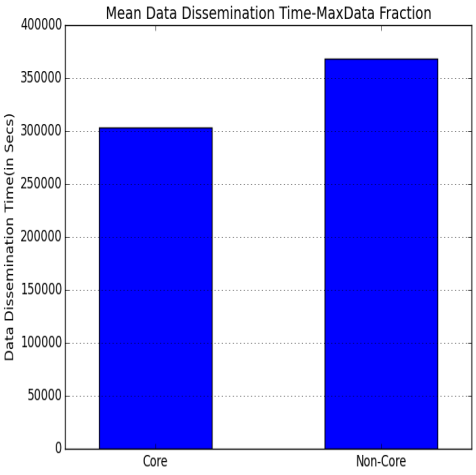

(b) PERCOM



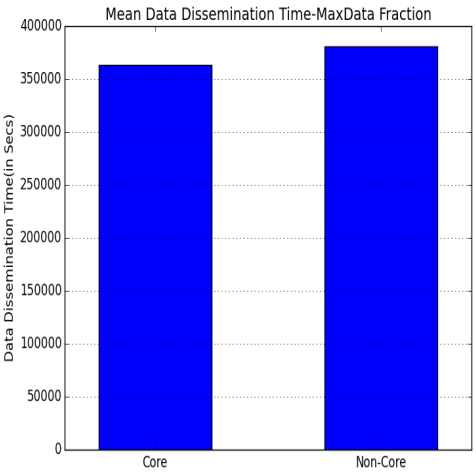

(c) ROLLERNET




(d) MIT




(e) MACACO-France




(f) MACACO-Brazil

Figure 7.7. Mean data dissemination time obtained from different traces using Core (Best Relays) and Non-Core nodes for maximum data fraction.

group of students and researchers in the university. From Figure 7.7 depicts the effectiveness of *Best Relays* obtained from *BROP Component*. With the help of Core nodes (or Best Relays) as initial data sources, the data disseminates faster in the network. As evident from the Figure, data dissemination time with Core nodes as starting points is much lower compared to the data dissemination time when initiating from Non-Core nodes. We also see that, the more the network is loosely connected, the higher is the gain. For ROLLERNET trace, we do not see much difference due to closed knit structure where each node is strongly connected to others in the network. From the above results, we clearly observe the effectiveness of using the Core Nodes, and thus the importance of using the weighted K-Shell decomposition algorithm in *BROP Component*.

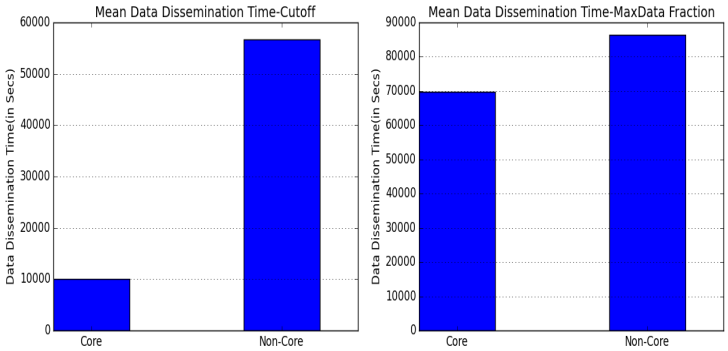
## 7.4 Results and discussion

After demonstrating the effectiveness of the Core Nodes usage, I integrate the *BROP Component* into INDIGO, and I quantify the advantage of using it in the case of data dissemination time till Cut-off point $\alpha$ and max amount of data $DF_{max}$, and I will discuss the reasons behind the different behavior.

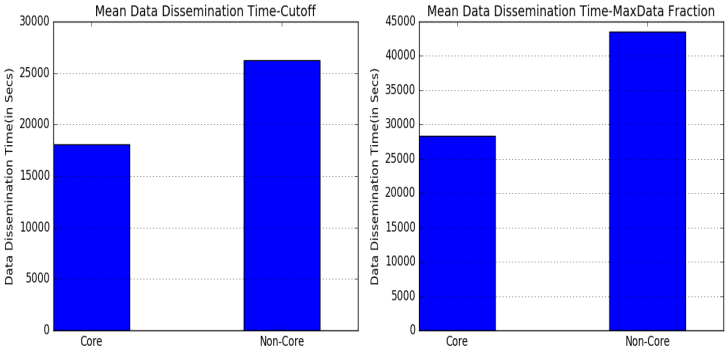
Figure 7.8, Figure 7.9, and Figure 7.10 present the impact of *Best Relays* on data dissemination time for different contact traces under the broadcast strategy.

Under the broadcast strategy, the *Best Relays* are obtained only by considering the contact patterns. In general, we observe that *Best Relays* are very effective in reducing the data dissemination time till the Cut-off point as compared to the time obtained through Non-Core nodes. This happens because when *BROP Component* starts data dissemination with *Best Relays* (or Core Nodes), they quickly disseminate information to their bigger ego-network and rapidly reaches to the Cut-off point data fraction. However, the difference in data dissemination time using Core and Non-Core reduces when I utilize them for $DF_{max}$. This happens because, in the case of $DF_{max}$ data requirement, *BROP Component* also need to consider the time required to diffuse information to the Non-Core nodes which lie in the outermost shell of the network (or not well-connected to rest of the network).

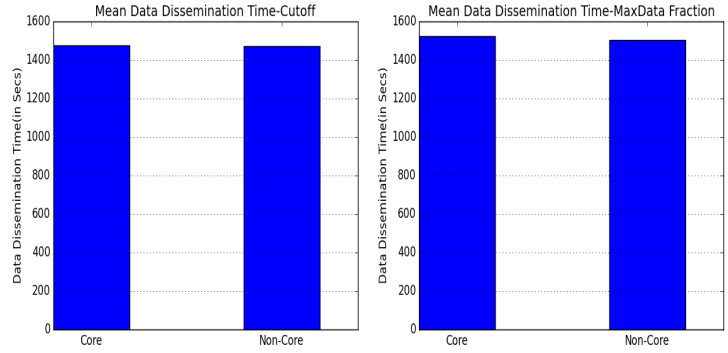
For both conference environment traces, we observe the same trend as described above. However, for urban environment ROLLETNET traces, we barely observe the impact of *Best Relays* due to its close-knit network structure where most of the people are very well connected to each other and lie in the core of the network. Figure 7.9 also present the similar pattern and impact of *Best Relays* on all weeks of MIT traces i.e. MIT-W1 to MIT-W4. For the first 2 weeks (MIT-W1




(a) INFOCOM

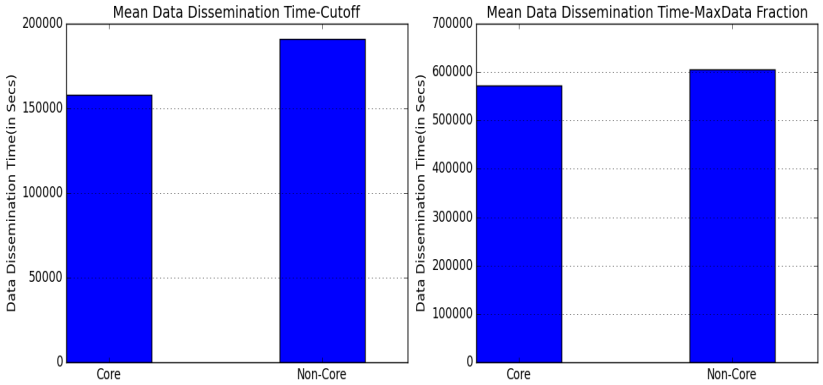



(b) PERCOM

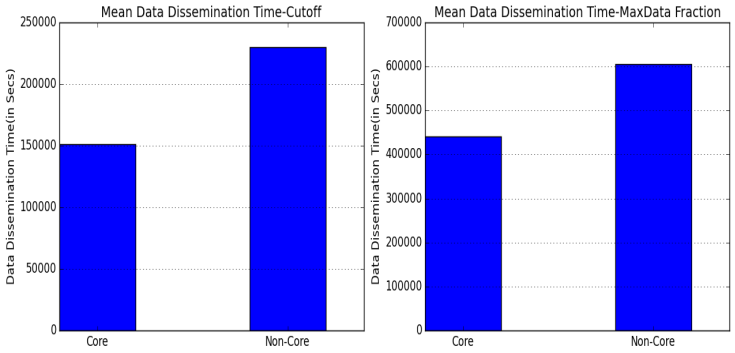



(c) ROLLERNET

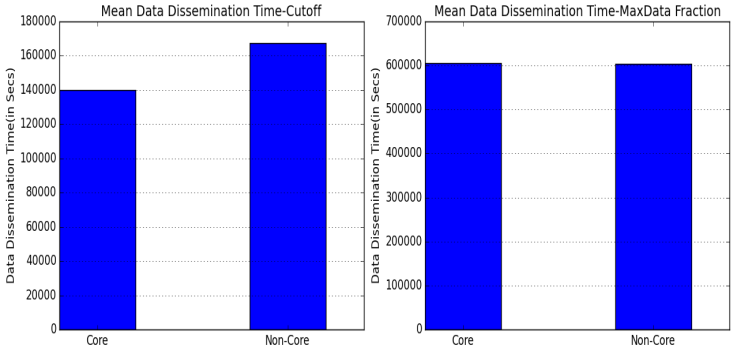
Figure 7.8. Mean data dissemination time obtained from different traces using Core (Best Relays) and Non-Core nodes for Cut-off point and maximum data fraction for INFOCOM, PERCOM, and ROLLERNET under the broadcast strategy.



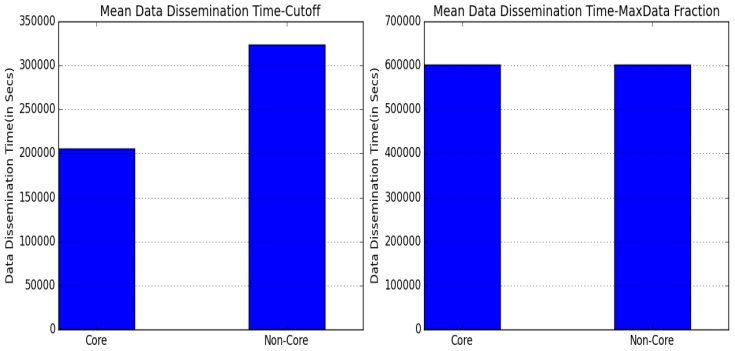

(a) MIT-W1




(b) MIT-W2

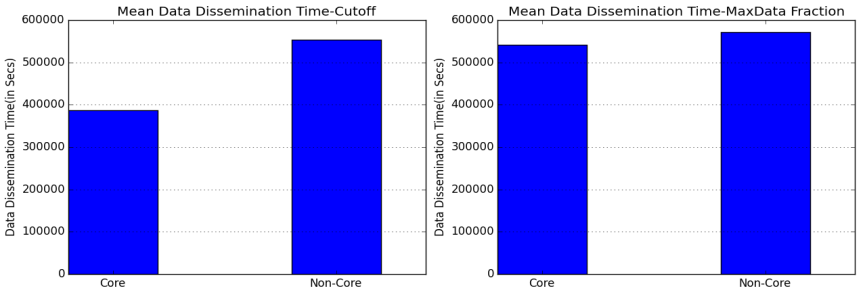



(c) MIT-W3

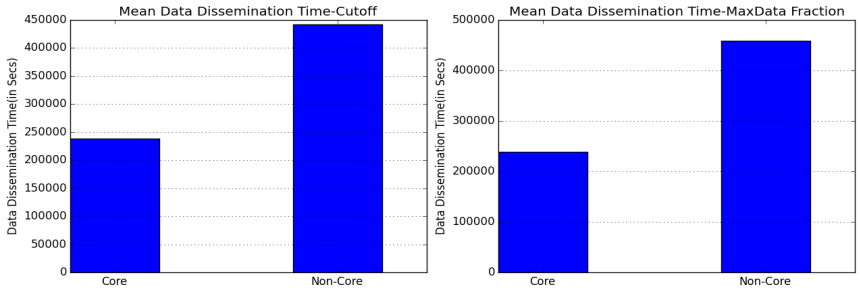



(d) MIT-W4

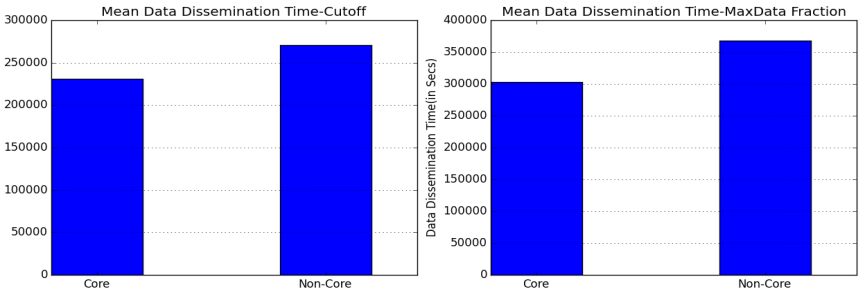
Figure 7.9. Mean data dissemination time obtained from different traces using Core (Best Relays) and Non-Core nodes for Cut-off point and maximum data fraction for all four weeks of MIT under the broadcast strategy.



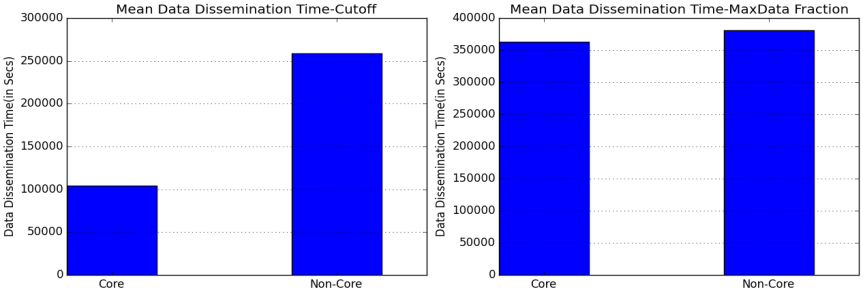

(a) MACACO-France-W1



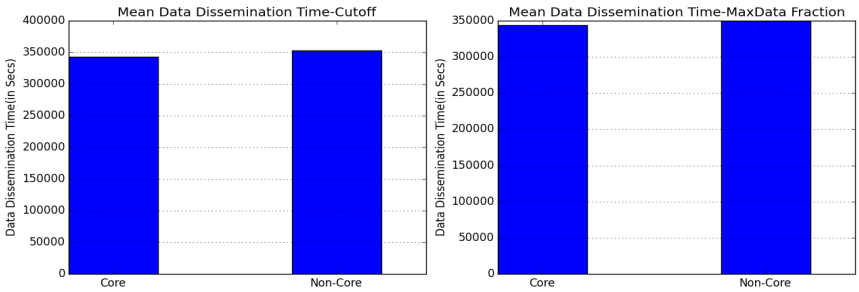

(b) MACACO-France-W2




(c) MACACO-France-W3




(d) MACACO-Brazil-W1



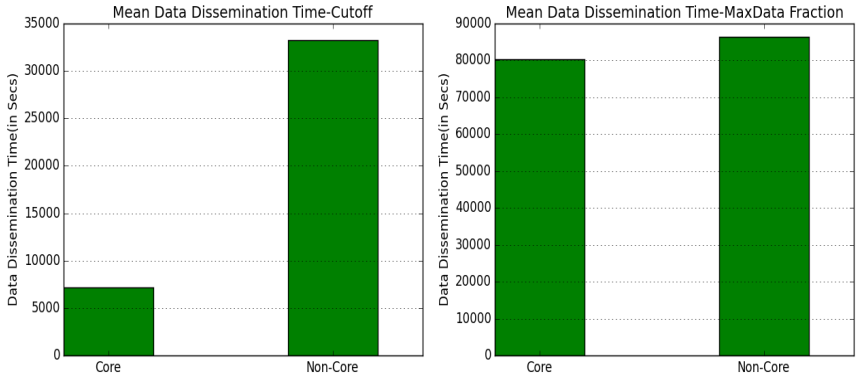

(e) MACACO-Brazil-W2

Figure 7.10. Mean data dissemination time obtained from different traces using Core (Best Relays) and Non-Core nodes for Cut-off point and maximum data fraction for all weeks of MACACO-France and MACACO-Brazil under the broadcast strategy.

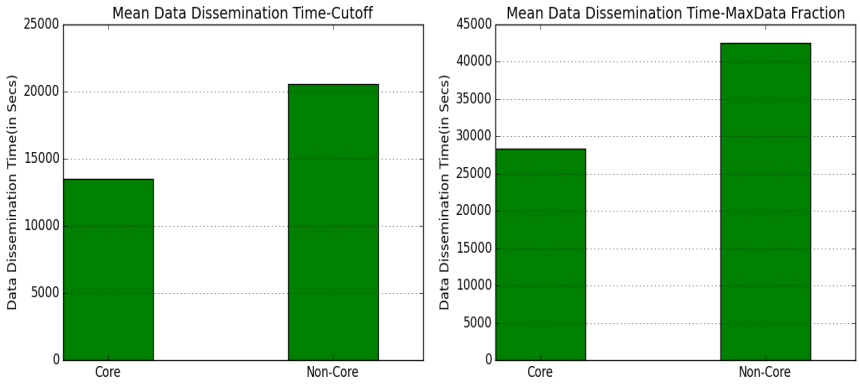
to MIT-W2), *Best Relays* fastens the data dissemination process by decreasing the data dissemination time. However, for MIT-W3 and MIT-W4 week, due to fewer contacts among people (semester break at MIT) we see the limited impact of *Best Relays* for $DF_{max}$. The similar trend is also observed in both groups of MACACO trace (from Figure 7.10) of the university environment. From above discussion, I find that *BROP Component* is quite effective in reducing the data dissemination time under the broadcast strategy by suggesting a set of *Best Relays* in the network.

The effectiveness of *BROP Component* is also evident under interest-driven data dissemination strategy from Figure 7.11, Figure 7.12 and Figure 7.13. Similar to broadcast strategy, once again the *Best Relays* obtained under interest-driven strategy is more effective for data fraction until Cut-off point as compared to the maximum fraction of data $DF_{max}$. While analyzing the Core and Non-core nodes achieved from K-Shell decomposition algorithm under interest-driven strategy, I also found the importance of blending of contact and interest similarity weights. More specifically, I observed the shifting of Core and Non-core nodes in different shells based on their interest-similarity. I also observed the change in network and found that interest-driven data dissemination strategy brings those nodes closer to the core-shell who are more central to the network according to their interests.

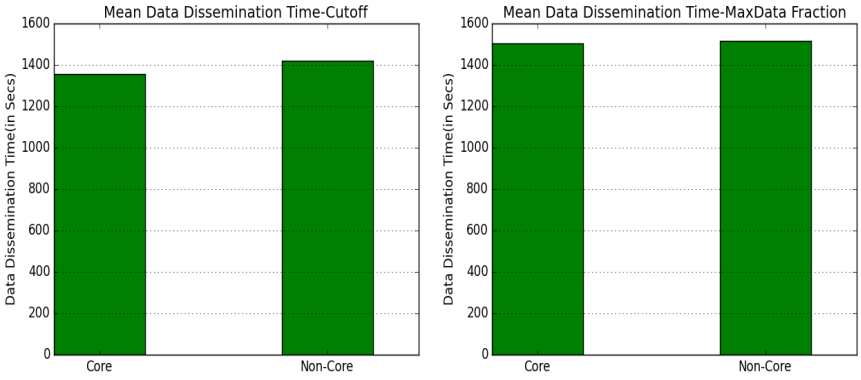
The results obtained for conference environment show pattern similar to the broadcast strategy. We also observe that for closed knit community like ROLLERNET, the shifting of nodes to the core is not evident thus, does not affect the data dissemination time. For MIT traces, we also observe a similar pattern, however, for MIT-W2 we observe that for max data fraction, the difference between the data dissemination time of Core and Non-core shortens. This might occur because the selected Core nodes are more central to the social proximity but might not exhibit very high contacts with other nodes of outer shells thus leads to higher data dissemination time. MACACO traces for both groups also exhibit the pattern similar to broadcast strategy. Finally, for all traces, I also observed that the mean fraction of data collected by *Best Relays* is higher ( 20%) than the fraction of data collected by Non-Core nodes.




(a) INFOCOM

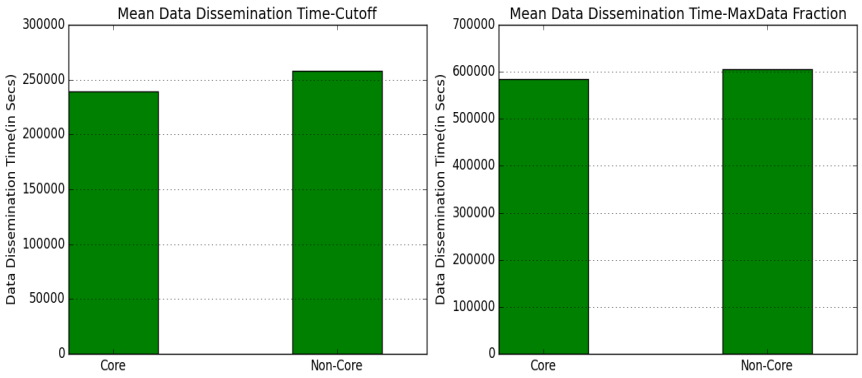



(b) PERCOM

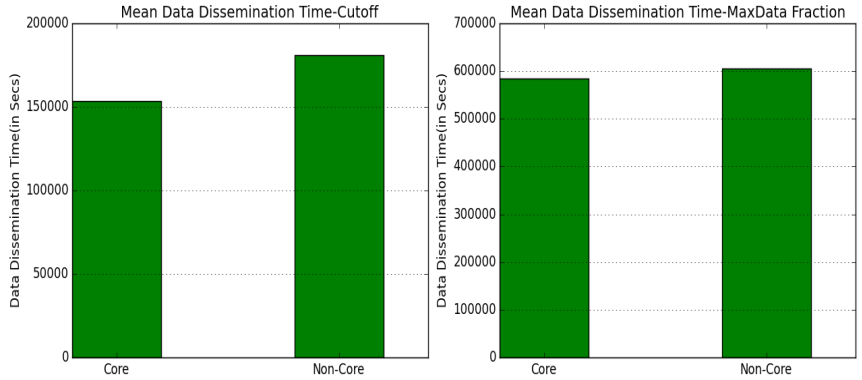



(c) ROLLERNET

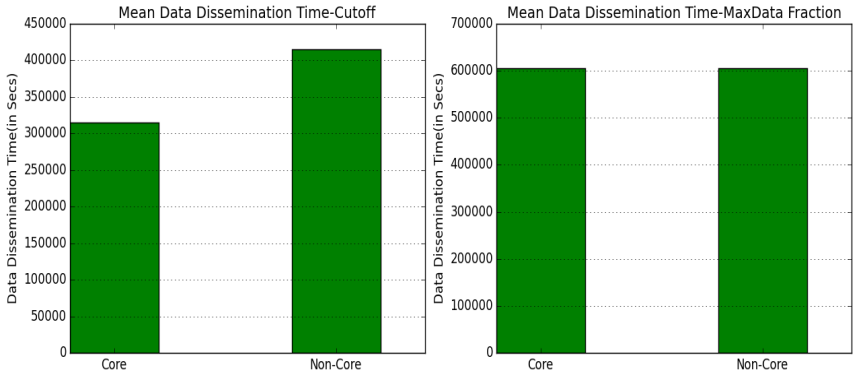
Figure 7.11. Mean data dissemination time obtained from different traces using Core (Best Relays) and Non-Core nodes for Cut-off point and maximum data fraction for INFOCOM, PERCOM, and ROLLERNET under the interest-driven strategy.



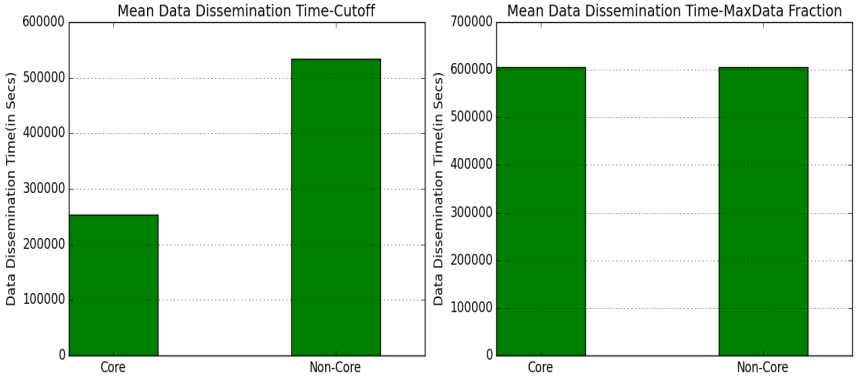

(a) MIT-W1




(b) MIT-W2

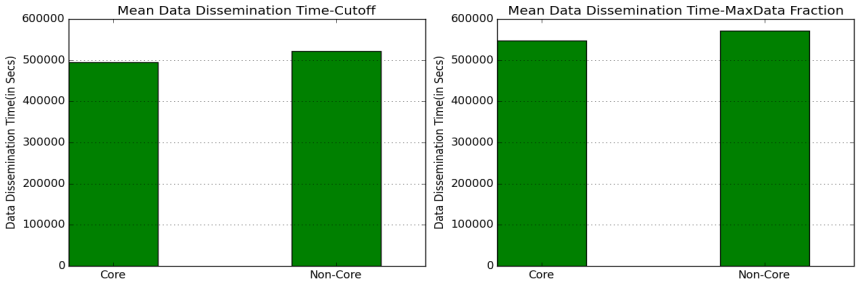



(c) MIT-W3

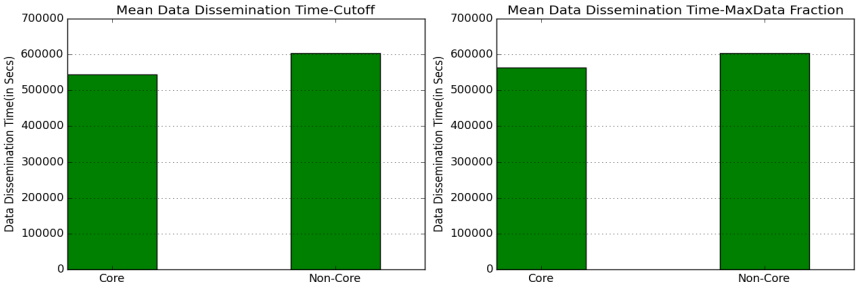



(d) MIT-W4

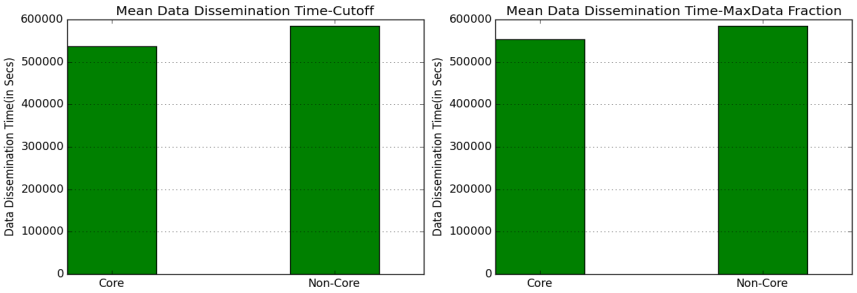
Figure 7.12. Mean data dissemination time obtained from different traces using Core (Best Relays) and Non-Core nodes for Cut-off point and maximum data fraction for all four weeks of MIT under the interest-driven strategy.



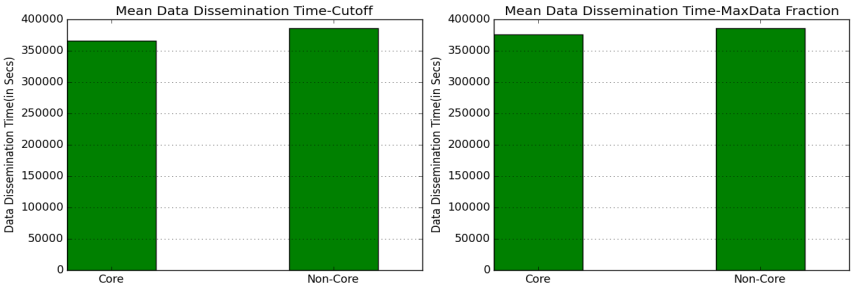

(a) MACACO-France-W1



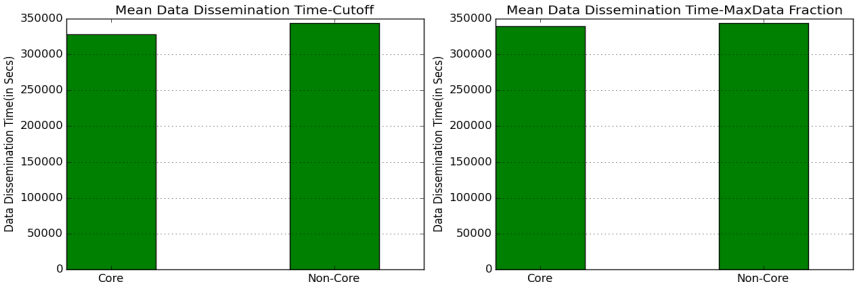

(b) MACACO-France-W2




(c) MACACO-France-W2




(d) MACACO-Brazil-W1




(e) MACACO-Brazil-W2

Figure 7.13. Mean data dissemination time obtained from different traces using Core (Best Relays) and Non-Core nodes for Cut-off point and maximum data fraction for all weeks of MACACO-France and MACACO-Brazil under the interest-driven strategy.

## 7.5 Conclusions

In this Chapter, I presented the estimation of *Best Relays* in the network using the *BROP Component* of INDIGO for both Type II and III cases. The methodology I proposed to find the *Best Relays* for broadcast and interest-driven data dissemination strategy uses the K-Shell decomposition algorithm that considers both degree centrality and links weights of each node in the network. The results obtained for *Best Relays* validates the usefulness of *BROP Component* for both data dissemination strategy. I summarize the key observation from this Chapter as follows:

- K-Shell decomposition algorithm adopted by *BROP Component* is useful for both broadcast and interest-driven data dissemination strategy and finding the set of *Best Relays* in the network. I also proposed a methodology to combine both physical and social weights to detect *Best Relays*.
- The utilization of *Best Relays* is very effective in reducing the data dissemination time till the Cut-off point.
- For the close-knit community like ROLLETNET, *Best Relays* does not add much value in faster information dissemination because most of the people are very well connected to each other.
- Interest-driven strategy changes the network structure through the shifting of Core and Non-core nodes in different shells. By giving importance to social proximity, the nodes with lesser contact and higher interest similarity can also be one of the important nodes in the network.

In next Chapter, I will focus on the gray area (Type IV) of the physical–social proximity table where social proximity plays the important role. I will present my efforts to model data dissemination using online social networks using the *INGIDO–OSN* part of the INDIGO framework.

# Chapter 8

# Data Dissemination Under Online Social Proximity

## 8.1 Introduction

In previous Chapters, I presented the modeling and prediction of data dissemination process through INDIGO framework for Type II and III cases of Physical-Social proximity table where the presence of physical proximity is the fundamental requirement to disseminate information in physical networks (refer Figure 8.1).

In this Chapter, I will present the *INDIGO-OSN* part of the INDIGO framework and will present my efforts to model data dissemination under Type IV (or gray area) case where we only have availability of the social proximity information. The best example for Type IV scenario is the data dissemination in online social networks (OSNs) where the widespread use of social networking sites like Twitter and Facebook allow users to generate and share information anywhere and anytime thus, allows data dissemination only with the help of social proximity information. For such cases, even though people do not necessarily encounter each other but they can still share the information to their friends or followers.

The receiver of a message in such large scale networks has an option either to relay or forward it to his/her followers. In Twitter, this process is called retweeting and typically users retweet a message if they consider it interesting and worth sharing with others. A sequence of retweets along the network is called information cascade. Due to this process of sharing, a large amount of content is generated on Twitter and also opened the door for other new research directions in the field of information spreading, advertising, recommendations and social data mining. For example, online advertisers can use this information for effi-

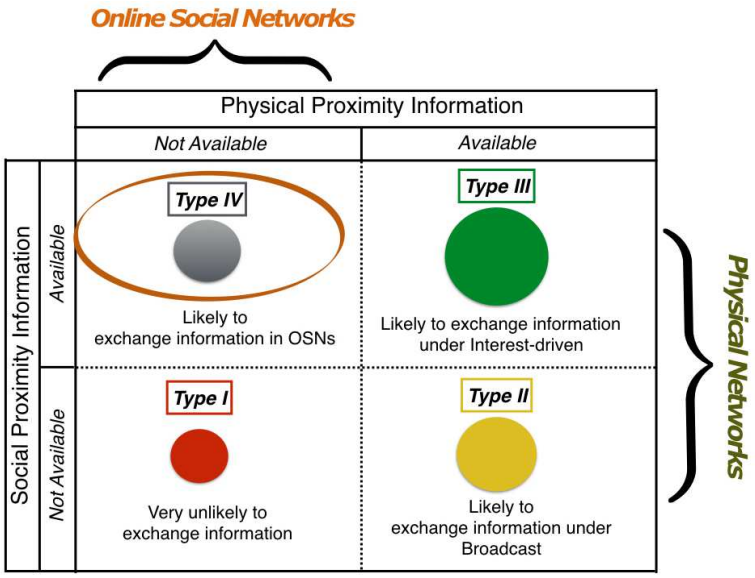



Figure 8.1. Type IV case of Physical-Social proximity table where the social proximity information among people is available. The exchange likelihood of Type IV is different from the one of Type II and III as in the first case the user has to choose to exchange, while in Type II and III it can happen without any action of the user due to an encounter.

cient targeted marketing campaigns. Media companies can learn how to effectively generate buzz for new films or shows. Political groups can also learn who they should try to influence in order to spread their message as far as possible.

The data dissemination in such scenario can be modeled by predicting the likelihood to retweet and reply to a given text (or tweet) by a particular user. As opposed to other works, I also take into account reply for information propagation by predicting the likelihood to reply to a particular tweet. Since a rich set of information is available on Twitter starting from users profile to content analysis, therefore, in this thesis, I not only provide a way to predict data dissemination by predicting the likelihood of retweet and reply but also quantify the importance of different information by introducing the concept of feature planes and model different real-world data dissemination aspects like heterogeneous activity of users on online social networks, type of information that needs to be disseminated, friendship ties and the content of the published online activities. Existing works in this area mainly tried to predict information spread by utilizing specific aspects of information like social network structure, temporal properties, profile features and topical features Galuba et al. [2010]; Petrovic

et al. [2011]; Yang and Counts [2010]; Pezzoni et al. [2013] but none of them successfully combined all these features together and, more importantly, they do not quantify the importance of different features for retweet prediction. I argue that a fundamental knowledge of different feature planes (defined as a group of features with similar cost in terms of privacy and complexity to acquire), their individual and combined contribution in retweet prediction has to be analyzed for better prediction of information diffusion. In this way, I not only address the problem of data dissemination for Type IV case but also enable to reduce the complexity of the model by providing the trade-off between high prediction accuracy and privacy. Also, the proposed model of INDIGO framework does not limit the prediction of retweet and reply to tweets that are generated by the friends of the user rather predicts it for any generic tweet.

In next sections, I will first describe the dataset used for Type IV data dissemination (Section 8.2). Further, I introduce the data dissemination approach using Machine Learning methods and the feature planes classification in Section 8.3. Section 8.4 presents the results obtained for different feature planes and validate my model. Finally, I conclude the Chapter in Section 8.5.

## 8.2 Dataset description

The dataset used in this work is the data collected from the Twitter activity of a large sample of about 2 Million users downloaded at IIT-CNR Pisa, 2013 Arnaboldi et al. [2013] presented in Figure 8.2. The dataset has been crawled through Twitter REST API [1], starting from a popular user in the network and then downloading all the available information about user's tweets and profile. Subsequently, the crawler iteratively downloaded same information for all the followers and friends of each user.

For each user in the dataset, the complete history of their tweets, retweets, and replies they posted on Twitter up are collected (up to the limit of 3,200 tweets per user imposed by Twitter REST API). In total, the dataset contains more than 2 Billion tweets, each of which is characterized by creation time, the id of the creator, textual content, the number of retweets it received, information about geolocation, and the set of entities it contains such as hashtags, ids of other users mentioned in the text, URLs, etc. In addition, each tweet also contains information about possible directed interactions between users. For retweets, this includes the id of the user who created the original tweet (i.e., the tweet that has been retweeted) and the creation time of the original tweet. For replies, the

[1] https://dev.twitter.com/rest/public

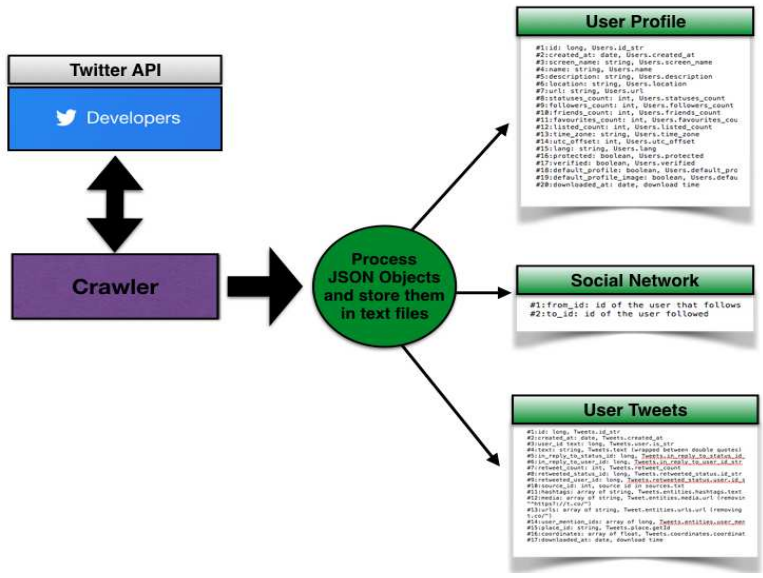
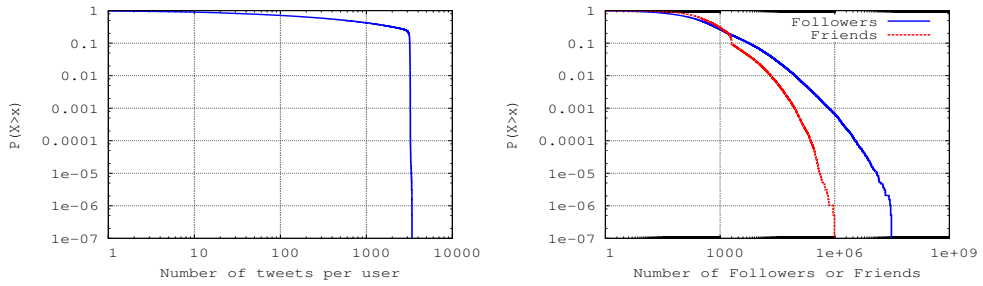



Figure 8.2. The process to collect Twitter data and the type of collected data.

tweet includes the id of the user who replied. The profile data downloaded for each user includes general user's information, such as user's name, description, geolocation, language, a personal URL, as well as some statistics about user's Twitter usage like a total number of tweets created, and the number of followers and friends.

Figure 8.3(a) depicts the CCDF of the number of tweets created by each user. It is worth noting that the distribution is truncated around 3,200[2] for the limit imposed by Twitter API. Nonetheless, the number of Twitter users who reach this limit are roughly 10% of the total number of users in the dataset. This means that for the majority of people the dataset has the complete history of tweets they created. In addition, for the users who created more than 3,200 tweets, the dataset has a large sample of their recent Tweeting history. Figure 8.3(b) depicts the CCDF of the number of followers and friends per user. Both graphs show a very long tail, with a very small fraction of users in the dataset reaching about one million of friends, and more than 20 million followers. This is a typical aspect of social networks and indicates the validity of our sample.

[2]For some users, the number of tweets is slightly larger than 3,200 since multiple downloads during the set-up process of the crawler were performed, which lasted roughly one month. Due to this for some users, there are additional tweets generated during this month.




(a) CCDF as a function of the number of tweets created per user

(b) CCDF as a function of the number of followers and friends per user.

Figure 8.3. Complementary cumulative distribution function of the number of tweets created and number of friends and followers per user

## 8.3 Data dissemination prediction methodology

Figure 8.4 presents the overall approach to predict data dissemination in online social networks using the *INDIGO–OSN* part of the INDIGO framework. From the collected Twitter data, I first clean and process the data to create features belonging to different planes based on their complexity to acquire and privacy intrusiveness[3]. Afterward, for each feature plane, I train and test the Machine Learning based multi-classification model to predict the data dissemination through retweet and reply. My approach allows a very effective single step data dissemination prediction and quantifies the influence of different feature planes on prediction results.

### 8.3.1 Data cleaning and processing

From the collected dataset, I first process the data to only consider the English tweets. To do this, I detect English language using the *langdetect* package of TextBlob [4] based on NTLK Language Toolkit [5]. Out of these English tweets, I create the ground truth to train the model by annotating them as the tweet, retweet, and reply where retweet and reply represent data dissemination while tweet signifies no-dissemination and call them type 0, 1 and 2 respectively. The data that needs to be disseminated is the text of the tweet. From processed data set with English tweets, I calculate features over time to capture possible

[3]The classification of features in different planes is provided according to our understanding as there is no formal classification of features according to privacy intrusiveness and complexity.

[4]textblob.readthedocs.io/en/dev/quickstart.html

[5]http://www.nltk.org/

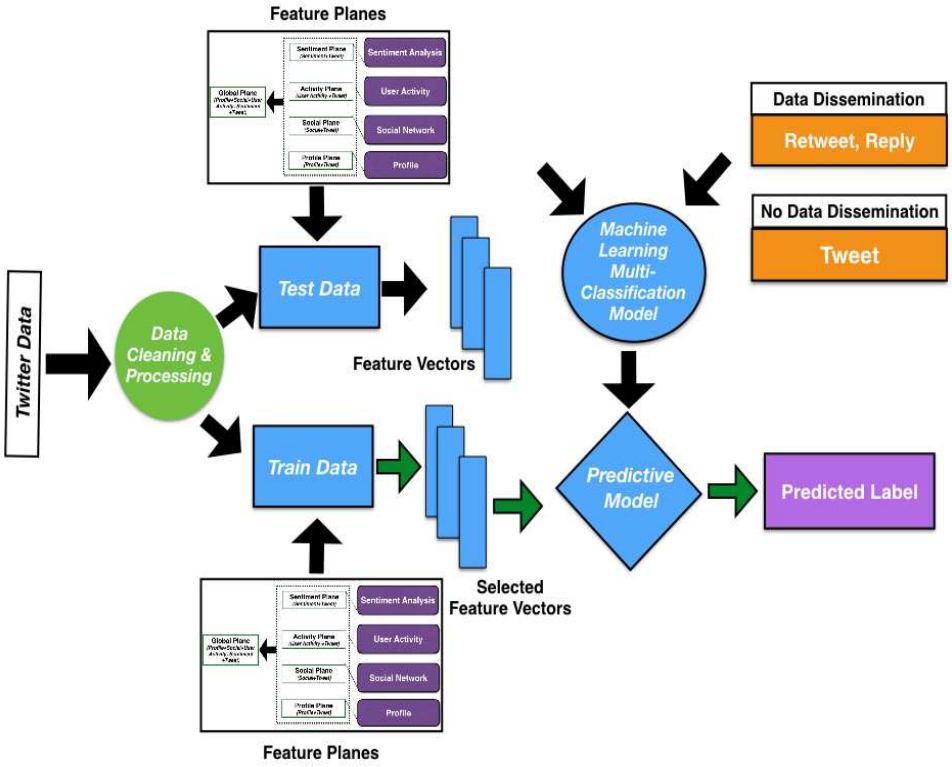



Figure 8.4. Methodology to predict data dissemination in online social networks using different feature planes and multi-classification prediction model.

changes in retweet and reply behaviors with time and generate a time series for each variable. Further, I aggregate these features in a weekly time window and store them in an SQLite database separately for each feature plane. The weekly aggregation was a good trade-off between precision and complexity because with the daily aggregation the complexity of the model will be too high for such a large scale data.

### 8.3.2 Feature Planes

To model a person's likelihood to retweet and reply, I propose different planes of features and extract them from Twitter data according to the increased complexity to acquire them and their privacy intrusiveness in INDIGO framework. From the privacy point of view, we consider how much information do we need to mine and reveal about a user in order to predict retweet and reply. The consideration of privacy during Twitter data mining is also highlighted in recent studies Kelley and Cranshaw [2013] Gan and Jenkins [2015]. Based on these contexts, I pro-

pose different feature planes starting from profile features to sentiment analysis of tweets. In this way, I also show the usefulness of rich information available in the dataset. Figure 8.5 presents different planes of features considered in the paper starting from *Profile* to *Global* plane. Please note that in each feature plane, I also consider features associated to the current Tweet. With *Tweet features*, I intend to examine the popularity of the original tweet and time sensitivity Petrovic et al. [2011]. Other Tweet features considered in each plane are the sentiment of the tweet, the number of embedded mentions and URLs obtained through the tweet inspection.

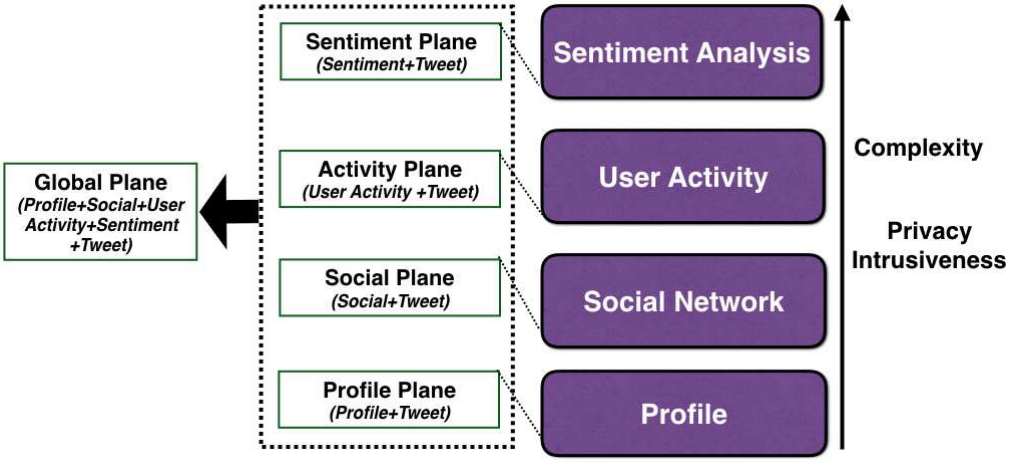



Figure 8.5. Feature planes based on the complexity to acquire and privacy intrusiveness starting from user profile feature to sentiment analysis of tweets.

- **Profile Plane:** Features associated with this plane are the easiest to acquire using public Twitter API[6]. From the Twitter profile of a user, I intend to get information about the user's account history like the length of user screen name, availability of URL, user description, and image on his/her profile. I hypothesize that users with longer account history and rich profile information may be more active on Twitter, therefore, it is more likely to predict their likelihood to retweet and reply. Additionally, I also capture social information of the user from their profile by extracting the number of friends, the number of followers and the number of groups a user is associated with (listed count). Finally, from user profile I also include the activity of users through their status counts (how many tweets users has published

[6]The profile data of a user can be accessed through a single Twitter API call.

recently) and favorite counts (how many tweets has been marked favorite by a user) features.

- **Social Plane:** Features in this plane represent the social ties of a person. Intuitively, if a person has more friends and followers then he/she has a higher probability to retweet and reply. Recent works also show that potential of retweeting as an act of friendship and to gain followers Boyd et al. [2010]. In this context, I process each user's network of friends and followers and extract features related to the number of friends, the number of followers, ratio of a number of friends to the number of followers and, ratio of a number of non-friends to the number of followers. As compared to *Profile* plane features, *Social* plane features are difficult to acquire and more privacy intrusive as we look into the entire social network of users.

- **Activity Plane:** This plane captures all past and recent activities of Twitter users to predict their willingness to retweet and reply. I assume that if a person exhibits more activity on Twitter, then it is more likely that he/she will retweet and reply. I also extract user's activity with respect to their friends, followers, and strangers like descriptive statistics for tweets per follower, friends, and strangers. *Activity* plane features are even more difficult to acquire and more privacy intrusive because in this case, we inspect all tweets of users to extract statistics about their past tweet, retweet and reply behavior with other users. In this plane, I capture both past and recent activities of users. For past activities, I utilize all available tweets up to current time while for recent activities I only take into account past month data (i.e. four weeks).

- **Sentiment Plane:** The features associated with this plane are the most computational costly and privacy intrusive as compared to other planes because, in this case, I inspect the content of each tweet and process them to find associated positive, negative or neutral sentiment. Similar to *Activity* plane features, I also extract all past and recent sentiments of tweets and also quantify tweet sentiments for friends, followers, and strangers. To measure the overall sentiment of a set of tweets (or retweets/replies) in a day, I define sentiment index $SI$ in Equation 8.1 where $s^+$ represents positive sentiment and $s^-$ presents negative sentiment values in a day. To calculate $SI$, I perform sentiment analysis only on English tweets from data set for each user and on day-wise tweets using TextBlob [7]. To calculate $SI$

[7] textblob.readthedocs.io/en/dev/quickstart.html

Table 8.1. Feature Set Input For Prediction Model

| Feature Plane | Feature Set |
|---|---|
| Profile | $< UserID, TweetFeatures, ProfileFeatures, TweetType >$ |
| Social | $< UserID, TweetFeatures, SocialFeatures, TweetType >$ |
| Activity | $< UserID, TweetFeatures, UserActivityFeatures, TweetType >$ |
| Sentiment | $< UserID, TweetFeatures, SentimentFeatures, TweetType >$ |
| Global | $< UserID, TweetFeatures, ProfileFeatures, SocialFeatures,$ $UserActivityFeatures, SentimentFeatures, TweetType >$ |

values, I only consider tweets whose sentiments can be classified through TextBlob library. Likewise, I calculate $SI$ values for each day of the tweets corresponding to each user.

$$SI = \frac{\sum s^+ - \sum s^-}{\sum s^+ + \sum s^-} \tag{8.1}$$

- **Global Plane:** This plane combines all features from *Profile*, *Social*, *Activity*, and *Sentiment* planes along with Tweet features. With the help of this plane, I intend to study the aggregated impact of all feature planes on data dissemination.

Utilizing the notion of feature planes, I create a final set of features for a given user and tweet pair $< u, tw >$ to train the prediction model. To create *Activity* and *Sentiment* plane features for $< u, tw >$ pair, I extract data only till the current time of the tweet $tw$. Please note that, since the features for *Profile* and *Social* planes do not change with time for a given user, they remain static for a given $< u, tw >$ pair. Table 8.1 presents the format of feature sets for all planes given as an input to train the prediction model.

### 8.3.3 Multi-classification prediction model

The prediction of data dissemination in OSNs can be done by predicting the likelihood of a person to retweet and reply a given text (or tweet). Therefore, this problem is a multi-classification problem where I need to predict retweet, reply (data dissemination occurs) and a tweet (if no data dissemination happens).

To do this, I employed several Machine Learning methods like Linear Regression Weisberg [2005], Random Forest Breiman [2001], Support Vector Machines (SVMs) Suykens and Vandewalle [1999] and Gradient Boosting Machine Friedman [2001]. Out of all these methods, Gradient Booting Machine outperforms for different samples of twitter data. Therefore, to predict data dissemination in social networks, I finally considered Gradient Boosting Machine method, a supervised machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models. More specifically, I use the XGBoost or "Extreme Gradient Boosting" method to classify retweet and reply. The detailed explanation of XGBoost is presented in Chapter 4.

The implementation of Gradient Boosting Method is based on the Python library XGBoost [8]. To classify retweet, and reply, I utilize multi-class classification using the *softmax* objective function. Further, I tried a set of parameter combinations to prevent overfitting using three parameters, *eta* that determines the learning rate, *gamma* regulating the sensitiveness to training examples, and the *number of rounds*. Based on different experiments, I set *eta* and *gamma* as 0.1 and 0 respectively. I also apply *10-fold cross-validation* to select an appropriate *number of rounds* based on the multi-classification error rate. For a given $< u, tw >$ pair, the prediction model predicts the likelihood of diffusion by classifying retweet, and reply. If the model predicts retweet and reply for a $< u, tw >$ pair then, the single-step diffusion will occur otherwise, there will be no diffusion. Differently, from existing solutions, my model enables retweet and reply prediction for any generic tweet and does not limit the prediction for tweets generated by someone connected to the user.

## 8.4 Results and discussion

To show the validity of the model of INDIGO framework and its usefulness of data dissemination in online social networks, I measure precision and recall obtained from XGBoost model for the retweet, and reply classification. I split the set of tweets into a training and a testing set based on the timestamp of the tweets. The training set consists 60% of all tweets and the remaining 40% of the data is used to evaluate the prediction quality. I tested the model on two different samples (Sample 1 and Sample 2) of dataset selected based on different time intervals with 673,858 and 1,031,116 tweets respectively. Sample 1 data consists only one-month tweets of users while Sample 2 have all tweets of users for all

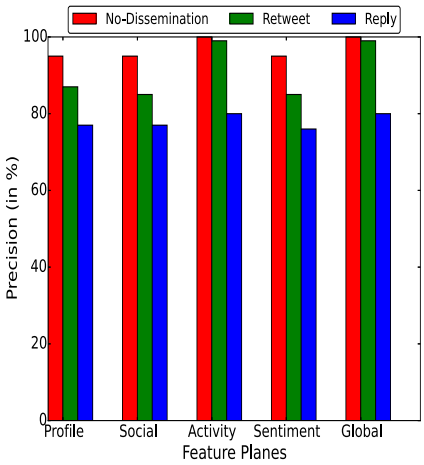
[8] *xgboost.readthedocs.io/en/latest/python/python_intro.html*

years. For each sample, I tested model accuracy for different planes of features starting from *Profile plane* to *Global plane*. Figure 8.6 presents the precision and recall obtained from both samples for all feature planes for the dissemination (retweet, and reply) and no-dissemination classification. Precision and recall are typical performance metrics used for measuring algorithm performance in Machine Learning. The Precision is defined as the number of true positives ($T_p$) over the number of true positives plus the number of false positives ($F_p$). The Recall is defined as the number of true positives ($T_p$) over the number of true positives plus the number of false negatives ($F_n$).

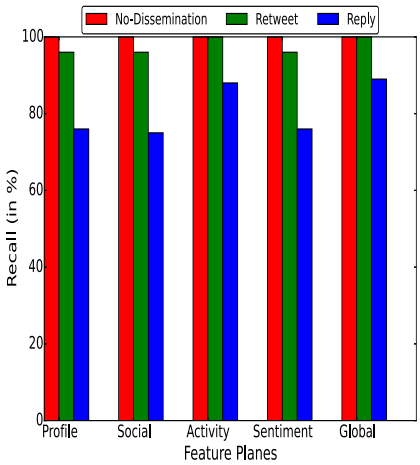
From results, we observe that for both samples, *Activity*, and *Global* plane features outperform and provide retweet, and reply classification with 99% and 82% precision and 99% and 80% recall values. Further, my model is also able to correctly classify no-diffusion with high precision (99%) and recall (100%) values. These results show that if we process and mine more information about users, the model becomes more precise in classifying retweet, and reply. We also observe that the model performs slightly better ( 2%) in *Profile* plane as compared to *Social* and *Sentiment* planes. This happens because, in *Profile* plane, we have more information about the user in terms of the number of status messages, association to groups while *Social* plane only has high-level information about friends and followers and *Sentiment* plane only considers sentiment of tweets. From above results, we observe the importance of the profiles of users and their activities on Twitter.

The precision and recall obtained using *Activity* and *Global* planes are equivalent and show that the maximum precision can be achieved only by considering user activities on Twitter i.e. *Activity* plane features. The inclusion of other feature planes such as *Profile*, *Social*, and *Sentiment* do not further improve prediction results. My results highlight that only with Profile plane features, we can already achieve very high accuracy, thus my approach can be used to reduce the complexity of large data processing and the privacy concerning issues. Finally, I also show the confusion matrix for both sample 1 and 2 in Figure 8.7 and 8.8 respectively. From both confusion matrix, we observe that the prediction model is able to correctly classify retweets, and reply as well as no diffusion for all planes thus, confirms the results obtained from Figure 8.6.

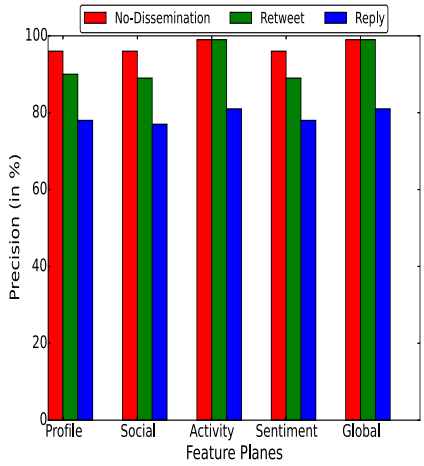
Table 8.2 presents the most important features associated with each plane utilized by our prediction model. From Table 8.2, we observe that Tweet features are one of the most important features across all planes. The tweet related features that contribute the most to precise prediction results are the time of the tweet, a number of times the tweet has been retweeted (Retweet count) and length & sentiment of the tweet. Since Tweet features are associated with each



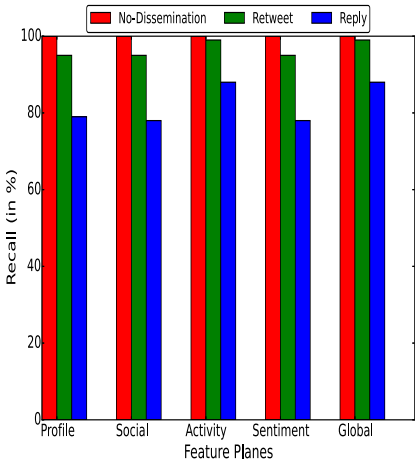

(a) Sample 1-Precision




(b) Sample 1-Recall




(c) Sample 2-Precision




(d) Sample 2-Recall

Figure 8.6. Precision and Recall obtained from different models utilizing different planes of features starting from Profile to Global plane.

plane, therefore, we also quantify their impact on model accuracy and observe that they contribute 30% to the overall model accuracy across all planes. Our prediction model obtains similar results for both samples (sample 1 one month data while sample 2 with years of data) across all planes. Therefore, our results show that only with one-month of the Twitter activity for a set of users is enough for accurate predictions. This result provides the implications for the amount of data required for the tweet, retweet, and reply classification and could be utilized in future diffusion models.


(a) Profile Plane (b) Social Plane (c) Activity Plane

(d) Sentiment Plane (e) Global Plane

Figure 8.7. Confusion Matrix for Tweet, Retweet, and Reply classification obtained from our model utilizing different feature planes starting from Profile to Global for Sample 1.

Compared to other resharing prediction models in the literature, we obtain sensibly higher accuracy values. For example, the model presented in Hong et al. [2012], which is, to the best of our knowledge, the only model that can be directly compared to ours, obtains prediction accuracy lower than 80%. It is also worth noting that this model limits the prediction to tweets only generated by friends of the target users, whereas in our model we calculate the likelihood to retweet or reply a generic tweet, not necessarily generated by someone connected to the selected user.

Finally, I also validate the applicability of INDIGO's prediction model for different time periods. To do this, I further group our testing data in the order of time (hour, day and week) after the last tweet of training data. For example, in the case of one hour, I only classify tweets that have been generated at max one hour after the last tweet in training data. Similarly, for days and week, I only classify those tweets that have been generated till the current day or week. From the results, we observe that for testing tweets generated up to one day after the last tweet of training data, the model predicts data dissemination with slightly higher precision ( 2%) for all planes except *Activity* and *Global* planes. In the case of *Activity* and *Global* planes, the precision obtained from the model was


(a) Profile Plane (b) Social Plane (c) Activity Plane

(d) Sentiment Plane (e) Global Plane

Figure 8.8. Confusion Matrix for Tweet, Retweet, and Reply classification obtained from our model utilizing different feature planes starting from Profile to Global for Sample 2.

same across different time periods thus, show the preciseness and applicability of the model for different time periods and makes my model time independent. This happens mainly due to our rich dataset and consideration of both recent and overall past activities of Twitter users and the right features selected from the Gradient Boosting model.

## 8.5 Conclusions

In this Chapter, I presented a novel approach to predict the data dissemination for Type IV of the Physical-Social table where social proximity plays a fundamental role using the *INDIGO–OSN* part of the INDIGO framework. I presented a Gradient Boosting Machine based multi-classification model to predict the data dissemination by predicting the likelihood to retweet, and reply. I also provide the deeper understanding of the diffusion process and quantifies the impact of the rich set of information available on online social networks by introducing feature planes: *Profile*, *Social*, *Activity*, *Sentiment*, and *Global* based on the complexity to acquire and privacy intrusiveness. I validated the proposed model on

Table 8.2. Important Features For Different Planes

| | |
|---|---|
| Profile Plane | Tweet time, # Followers, Tweet length, Twitter account age, # status messages, # Friends, Retweet count, Tweet sentiment, Listed Count, Length of user description |
| Social Plane | Tweet time, Ratio of friends and followers, # Friends, Tweet length, # Followers, Retweet count,Tweet sentiment |
| Activity Plane | Retweet count, Tweet length, Time elapsed since last Retweet, Tweet time, # Mentions, Tweet sentiment, STD of inter Retweet time, STD of # urls in Retweet, # Hashtags, Min. of inter Reply time, Mean of inter Tweet time, Time elapsed since last Tweet, Max. of total Retweets per follower, # Url |
| Sentiment Plane | Tweet time, Tweet length, Retweet count, Tweet sentiment, STD Retweet SI per follower, STD of Retweet SI, Max. of Retweet SI, Max. Retweet SI per follower, STD of Reply SI, STD of Tweet S, Entropy of Retweet SI, STD of Tweet SI per week,Entropy of Retweet SI |
| Global Plane | Retweet count, Tweet length, Time elapsed since last Retweet, Tweet time, # Mentions, Tweet sentiment, STD of inter Retweet time, STD of # urls in Retweet, # Hashtags, Min. of inter Reply time, Mean of inter Tweet time, Time elapsed since last Tweet, Max. of total Retweets per follower, # Url |

two different samples of the large-scale Twitter dataset and observe that it outperforms existing works for all planes by providing higher precision and recall for both samples. I also summarize the key contributions of this Chapter as follows:

- Proposed a model to provide a deep understanding of data dissemination by predicting the likelihood to retweet and reply and exploiting the different set of features by introducing feature planes. To the best of my knowledge, it is the first work that deeply studies the importance and impact of different planes of features on retweet and reply prediction.
- I define different planes of features that differ in complexity to acquire and level of privacy required.
- Differently, from existing solutions, my model enables data dissemination prediction for any generic tweet and does not limit the prediction for tweets generated by someone connected to the user.
- Results show high precision for data dissemination prediction for different planes and also present that user twitter activities feature plane provides the highest precision. Further, the results presented in this Chapter are also

seminal to researchers by providing the trade-off between high prediction accuracy and privacy.

The work done in this Chapter is published at "Complex Networks Workshop, 2015" and is also been selected to submit the extended version of the paper in Computation Social Networks Journal. This Chapter completes the modeling and prediction of data dissemination for all cases (starting from Type II to IV) of the Physical-Social proximity table and shows the effectiveness of the INDIGO framework solutions across all cases.

## 8.6 Remarks

The work done in this Chapter is done in collaboration with Valerio Arnaboldi of IIT-CNR, Pisa.

# Chapter 9

# Conclusions and Outlook

The main contribution of this thesis is to provide a set of solution to predict the performance of data dissemination by collectively considering the real-world aspects of data dissemination process under different based on the availability of physical and social proximity information among people. The solutions proposed in this thesis empowers local businesses or publishers to assess the performance of their localized dissemination based services in advance both in physical as well as the online social world. Furthermore, it relieves users from receiving large amounts of unnecessary data when there is no proximity. As shown in Figure 9.1 this thesis provides a solution called *INDIGO–Physical* for the cases where physical proximity plays the fundamental role (Type II and Type III). It enables the tighter prediction of data dissemination time and prediction of best relays under real-world mobility, communication, and data dissemination strategy aspects. Further, this thesis also contributes in providing the performance prediction of data dissemination in large-scale online social networks where the social proximity is prominent (Type IV) using *INDIGO–OSN* part of the INDIGO framework.

## 9.1 Summary and conclusions

The research work done in thesis contributed in the following directions.

### 9.1.1 Modeling and prediction of tighter upper bound of data dissemination time under real-world aspects

The first work done in this thesis provides a model that incorporates real-world mobility, communication, and dissemination strategy aspects for the tighter pre-

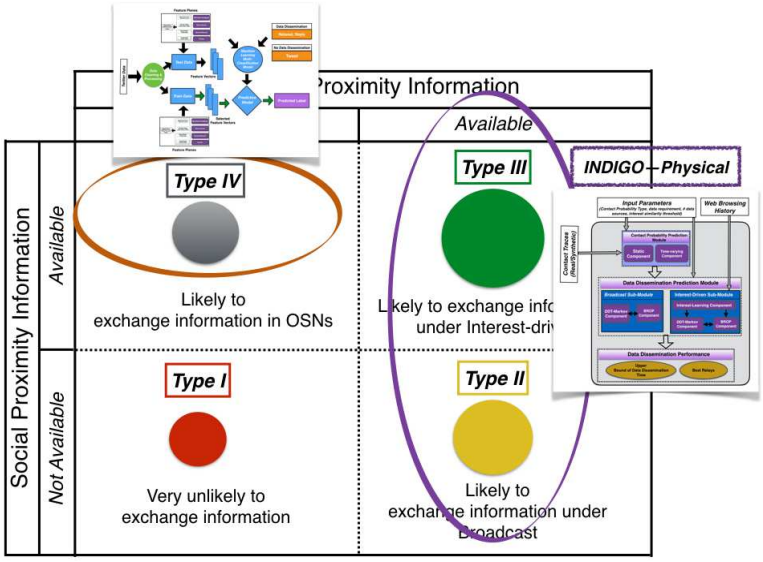



Figure 9.1. Solutions provided for Type II, III and IV of Physical–Social proximity table using different parts of INDIGO framework. No solution is required for Type-I as no exchange of information will happen in this case.

diction of data dissemination time for the cases where physical proximity plays the fundamental role. This thesis contributes in finding these aspects through the inspection of several real-world contact traces. The real-world aspects considered in this thesis are heterogeneous contact patterns (both static and time-varying), multiple simultaneous contacts among people, broadcast as well as interest-driven data dissemination strategy, various data requirements, and multiple data sources. Further, this thesis finds a long tail cut-off pattern in data gathering process and based on this property it develops a Markov chain based model called *DDT-Markov* that predicts the tighter upper bound of data dissemination time using a Cut-off point based approach. *DDT-Markov* is able to incorporate different aspects of data dissemination process under both broadcast and interest-driven data dissemination. The model is validated on 5 datasets (INFOCOM, PERCOM, ROLLERNET, MIT, MACACO) from 3 diverse environments (conference, urban area, university). It is able to achieve the tighter upper bound of data dissemination time within 5-15% error against the ground truth for both broadcast and interest-driven data dissemination strategy. The main observation for this part of the thesis is:

*"The data dissemination process exhibits a long tail cut-off pattern thus, we can identify the Cut-off point. The exploitation of such Cut-off point greatly contributes*

*towards the tighter prediction of the upper bound of data dissemination time"*

### 9.1.2 Prediction of heterogeneous time-varying contact patterns

The second novel contribution of this thesis is to provide a Gradient Boosting Machine Learning based model to enable the prediction of the time-varying pair-wise heterogeneous contact probabilities among people by capturing features related to their contacts and inter-contact time. The model learns the contact patterns of people and predicts their future day-wise contact probabilities. The model was validated on all datasets and achieves reasonable accuracy for contact datasets with longer duration. This happens because in long-term people exhibit regular contact patterns thus helps the model to predict accurately based on these patterns. The contact probability prediction accuracy increases as soon as we provided more learning data to the model. To get reasonable accuracy in contact probability prediction, the model requires at least one week of data. The main observation for this part of the thesis is:

*"A supervised machine learning approach can enable the prediction of future time-varying contact probabilities by learning the past contact patterns of people."*

*"Inclusion of time-varying contact probabilities further tightens the upper bound of data dissemination time."*

### 9.1.3 Learning of user interests

The third contribution of this thesis to provide a tool to learn the interests of users from their on mobile web-browsing history. This tool enables the automatic learning of user interests by finding the semantic categories of the websites using Term Frequency–Inverse Document Frequency (TF-IDF) weighting scheme. Using these interests, the tool also finds the interests similarities among each pair of people. Further, this tool also creates synthetic interests of people for datasets that do not capture web browsing history by utilizing the concept of strong medium and weak ties computed with a power law distribution. The main observation for this part of the thesis is:

*"The interest learning tool also produces interests specific to the country language."*

*"Inclusion of interest similarities among people enables the interest-driven data dissemination strategy in predicting the upper bound of data dissemination time and shows that interest-driven data dissemination strategies are effective to restrict the spread of dissemination as opposed to broadcast strategy because information is only disseminated to people with sharing similar interests."*

### 9.1.4 Collection of traces with mobility and interests

The fourth contribution of this thesis is collect the dataset that contains both contacts and interests of people. To the best of our knowledge, there were no real-world traces that contain this type of data. This thesis fills this gap by collecting data by utilizing a dedicated mobile application developed as a part of CHIST-ERA MACACO project in collaboration with project partners in France and Brazil. The application is also available at Google Play Store. The application collect information from different sensors like GPS, accelerometer, Wi-Fi scanning along with the on-mobile web browsing history. From this data, I used the Wi-Fi scanning data to create contacts among people (physical proximity) and browsing history of people to reflect their interests (social proximity).

*"This thesis collects and produces a dataset that captures both physical and social proximity among people."*

### 9.1.5 Prediction of best relays for faster data dissemination

The fifth contribution of this thesis is to find the best relays in the network to maximize the information diffusion or to minimize the data dissemination time. This thesis proposed a methodology to find the *Best Relays* for both broadcast and interest-driven data dissemination strategy by employing a weighted K-Shell decomposition algorithm that considers both degree centrality and links weights of each node in the network. The nodes that reside in the Core of the network were considered as *Best Relays* while the nodes that are situated near the periphery (or Non-core) of the network were the worst connected nodes. The results showed the usefulness of *Best Relays* by reducing the data dissemination time for both dissemination strategies. More specifically, utilization of *Best Relays* is very effective in reducing the data dissemination time till the Cut-off point.

*"By giving importance to social proximity, the nodes with lesser contact and higher interest similarity can also be one of the important nodes in the network, therefore, under interest-driven strategy the network structure changes through the shifting of Core and Non-core nodes."*

*"For close-knit communities, Best Relays does not improve information dissemination because most of the people are very well connected to each other."*

All the above described contributions of this thesis are incorporated in the *INDIGO–Physical* part of the INDIGO data dissemination framework to provide a unified solution and to predict the performance of data dissemination for the upper bound of data dissemination time and detection of best relays under Type II and III cases of the Physical–Social proximity table.

### 9.1.6 Modeling and prediction of data dissemination in online social networks

The final contribution of this thesis is to provide the prediction of data dissemination in online social networks (Type IV) using the *INDIGO–OSN* part of the INDIGO data dissemination framework. To do this, I proposed a Machine Learning model to provide a deep understanding of data dissemination by predicting the likelihood to retweet and reply and exploiting the different set of rich features available in the Twitter dataset. This thesis also introduced and defined multiple planes of features that differ in complexity to acquire and level of privacy required. The proposed solution also enables the data dissemination prediction for any generic tweet and does not limit the prediction for tweets generated by someone connected to the user. The model was validated on a large scale Twitter dataset and results show high precision for data dissemination prediction for different planes and also show the highest precision for the plane that captures user's twitter activities.

*"Different feature planes and their accuracy in predicting data dissemination are seminal to researchers by providing the trade-off between high prediction accuracy and privacy."*

## 9.2 Directions for future research

I will now present the future research directions that can be addressed to extend the work presented in this dissertation.

### 9.2.1 User profiling and modeling of data dissemination using Location Based Social Networks (LBSNs)

The INDIGO framework can be extended to incorporate and model the LBSN datasets like FourSquare, Gowalla etc. These datasets typically have users check-in information in different Point of Interests. From these datasets, the physical proximity among people can be created based on the co-location check-in information. Similarly, the social proximity among people can be created by learning their interests according to the types of places they are visiting. For example, FourSquare has a category tree consisting 10 types of high-level categories like Food, Art, Travel, Shopping for each Point of Interest. With the help of this information, the profile of a user can be generated. In fact, I have already developed my own *User Profiler* for the LBSNs that creates a profile of users across these

10 categories. A user's profile represents the preferences or interests of the user from the perspective of semantic categories of visited places. Once we have both physical and social proximity information about the users then we can provide it to the INDIGO to predict the performance of data dissemination in LBSNs. This extension could be very crucial for the businesses as they can predict in advance about the usefulness of their offers among their customers and other interested people. Further, they can also target people based on their learned profiles.

### 9.2.2 Prediction of user interests from their personality traits

A further extension of my work could be to predict the interests of users based on their personality traits. In this way, we can classify a set of user interests according to their personality traits. This study could be useful to assign real interests to users rather than creating synthetic interests (in case user's web browsing history is not available). Also for this part, I have done an initial research during my internship at Telefónica Research and found that by capturing the temporal patterns of mobile usage, we can predict the personality of people and can also associate their interests with it.

### 9.2.3 Interests modeling on online social networks using knowledge graphs

The work done for the prediction of data dissemination in online social networks can be further extended by introducing another plane called interests plane where the interests of Twitter users can be developed by exploiting the topics from their tweets using Knowledge Graphs. Using these graphs of topics, we can find the similarity between two users graphs or similarity between the tweet and a user's graph and use it as a feature to predict the data dissemination.

### 9.2.4 Prediction of complete cascades in online social networks

Another challenging and big extension of the current thesis is the prediction of complete cascades i.e. the complete flow of information starting from the originator. The work presented in this thesis only predicts the single step information diffusion, however, the model proposed in this thesis can be further extended to predict the complete cascades.

### 9.2.5 Usage of INDIGO for Internet of Things

Finally, another extension of INDIGO is to utilize it for Internet of Things (IoT) where data sources are IoT devices and humans can be the receiver of the information originating from these devices. INDIGO can be useful to model and understand the behavior of people in such scenario. Some findings of the work done in this thesis have also lead to receiving a grant of a CHIST-ERA European project called UPRISE-IoT[1].

The aim of the UPRISE-IoT project is to develop the U-HIDE solution which will empower the users to understand and make their own decisions regarding their data, which is essential in gaining informed consent and in ensuring the take-up of IoT technologies. This will include behavioral models that are building on the ones developed in the INDIGO framework, to allow profiling the privacy primitives, mechanisms, and techniques for data dissemination of the IoT devices.

[1] http://www.chistera.eu/projects/uprise-iot

# Publications

## 9.3 Peer-reviewed Articles

1. K. Garg, S. Giordano and M. Jazayeri "INDIGO: Interest-Driven Data Dissemination Framework For Mobile Networks". (Under Submission at IEEE WoWMoM 2017)

2. K. Garg, V. Arnaboldi and S. Giordano "A Novel Approach to Predict Retweets and Replies Based on Different Privacy and Complex-Aware Feature Planes", In Proceedings of The 5th International Workshop of Complex Networks and their Applications, November 2016, Milan, Italy.

3. S. Vanini, D. Gallucci, K. Garg, S. Giordano, V. Mirata and M. Bettoni "Modeling Social Interactions in Real Work Environments", In Proceedings of The 6th International Workshop HotPlanet 2015 in conjunction with ACM MobiCom 2015, September 2015, Paris, France.

4. K. Garg, S. Giordano and M. Jazayeri, "How Well Does Your Encounter-Based Application Disseminate Information?", In Proceedings of The 14th IFIP Annual Mediterranean Ad Hoc Networking Workshop (Med-Hoc-Net 2015), June 2015, Vilamoura, Algarve, Portugal.

5. K. Garg, S. Giordano and A. Förster, "A Study to Understand the Impact of Node Density on Data Dissemination Time in Opportunistic Networks", In Proceedings of 2nd ACM Workshop on High Performance Mobile Opportunistic Systems (HP-MOSys 2013), November 2013, Barcelona, Spain.

6. A. Förster, K. Garg, H. A. Nguyen, and S. Giordano, "On Context Awareness and Social Distance in Human Mobility Traces", in Third International Workshop on Mobile Opportunistic Networks (MobiOpp 2012), March 2012, Zurich, Switzerland.

## 9.4 Other Relevant Publications

### 9.4.1 Short Papers

1. Kamini Garg, "Data Dissemination Bounds in People-Centric Systems", In Proceedings of International Conference on Computer Communications (IEEE INFOCOM), April 2013, Turin, Italy (Student Workshop).

2. Kamini Garg, "COSN: A Collaborative Opportunistic Sensor Networking Framework for People-Centric Applications", In Proc. of The IEEE Pervasive Computing and Communication Conference (PerCom 2012), March 2012, Lugano, Switzerland (PhD Forum).

3. A. Hossmann-Picu , Z. Li, Z. Zhao, T. Braun, C. M. Angelopoulos, O. Evangelatos, J. Rolim, M. Papandrea, K. Garg, S. Giordano, A. C. Y. Tossou, C. Dimitrakakis and A. Mitrokotsa, "Synergistic User ⟷ Context Analytics", In Proceedings of The 7th ICT Innovations Conference, October 2015, Ohrid, R. Macedonia.

### 9.4.2 Poster and Demos

1. K. Garg, and S. Giordano, "Towards Developing a Generalized Modeling Framework for Data Dissemination", In Proceedings of the 12th European Conference on Wireless Sensor Networks (EWSN), Porto, Portugal, February 2015. **(Best Poster Award)**

2. A. Förster, K. Garg, M. Cabrini and S. Giordano, "Understanding and Optimizing Human Mobility with Smart Phones", In Proceedings of the 1st International Conference on ICT for Sustainability (ICT4S), Zurich, Switzerland, February 2013.

# Bibliography

T. Abdelzaher, Y. Anokwa, P. Boda, J. Burke, D. Estrin, L. Guibas, A. Kansal, S. Madden, and J. Reich. Mobiscopes for human spaces. In *IEEE Pervasive Computing, Vol. 6, No. 2*, pages 20–29. IEEE, 2007.

José Ignacio Alvarez-Hamelin, Luca Dall'Asta, Alain Barrat, and Alessandro Vespignani. k-core decomposition: A tool for the visualization of large scale networks. *arXiv preprint cs/0504107*, 2005.

Roy M Anderson, Robert M May, and B Anderson. *Infectious diseases of humans: dynamics and control*, volume 28. Wiley Online Library, 1992.

Valerio Arnaboldi, Marco Conti, Andrea Passarella, and Fabio Pezzoni. Ego networks in twitter: an experimental analysis. In *Proceedings of IEEE INFOCOM*, pages 3459–3464, 2013.

Pavlos Basaras, Dimitrios Katsaros, and Leandros Tassiulas. Detecting influential spreaders in complex, dynamic networks. *Computer*, 46(4):0024–29, 2013.

Vladimir Batagelj and Matjaž Zaveršnik. Fast algorithms for determining (generalized) core groups in social networks. *Advances in Data Analysis and Classification*, 5(2):129–145, 2011.

Farid Benbadis and Jeremie Leguay. CRAWDAD dataset upmc/rollernet (v. 2009-02-02). `http://crawdad.org/upmc/rollernet/20090202`, February 2009.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

Chiara Boldrini and Andrea Passarella. Data dissemination in opportunistic networks. *Mobile Ad Hoc Networking: Cutting Edge Directions, Second Edition*, pages 453–490, 2013.

Chiara Boldrini, Marco Conti, and Andrea Passarella. Contentplace: social-aware data dissemination in opportunistic networks. In *Proceedings of the 11th international symposium on Modeling, analysis and simulation of wireless and mobile systems*, pages 203–210. ACM, 2008.

Chiara Boldrini, Marco Conti, and Andrea Passarella. Design and performance evaluation of contentplace, a social-aware data dissemination system for opportunistic networks. *Comput. Netw.*, 54(4):589–604, March 2010. ISSN 1389-1286. doi: 10.1016/j.comnet.2009.09.001. URL `http://dx.doi.org/10.1016/j.comnet.2009.09.001`.

Chiara Boldrini, Marco Conti, and Andrea Passarella. Performance modelling of opportunistic forwarding under heterogenous mobility. *Computer Communications*, 48:56–70, 2014.

Phillip Bonacich. Power and centrality: A family of measures. *American journal of sociology*, pages 1170–1182, 1987.

Danah Boyd, Scott Golder, and Gilad Lotan. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In *System Sciences (HICSS), 2010 43rd Hawaii International Conference on*, pages 1–10. IEEE, 2010.

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

Han Cai and Do Young Eun. Crossing over the bounded domain: from exponential to power-law intermeeting time in mobile ad hoc networks. *IEEE/ACM Transactions on Networking (ToN)*, 17(5):1578–1591, 2009.

A. T. Campbell, S. B. Eisenman, N. D. Lane, E. Miluzzo, R. A. Peterson, H. Lu, X. Zheng, M. Musolesi, K. Fodor, and G. S. Ahn. The rise of people-centric sensing. In *Internet Computing, IEEE, 12(4)*, pages 12–21. IEEE, 2008.

Shai Carmi, Shlomo Havlin, Scott Kirkpatrick, Yuval Shavitt, and Eran Shir. A model of internet topology using k-shell decomposition. *Proceedings of the National Academy of Sciences*, 104(27):11150–11154, 2007.

Juan Miguel Carrascosa, Jakub Mikians, Ruben Cuevas, Vijay Erramilli, and Nikolaos Laoutaris. I always feel like somebody's watching me. measuring online behavioural advertising. *arXiv preprint arXiv:1411.5281*, 2014.

Arnaud Casteigts, Paola Flocchini, Walter Quattrociocchi, and Nicola Santoro. Time-varying graphs and dynamic networks. *International Journal of Parallel, Emergent and Distributed Systems*, 27(5):387–408, 2012.

Augustin Chaintreau, Pan Hui, Jon Crowcroft, Christophe Diot, Richard Gass, and James Scott. Impact of human mobility on opportunistic forwarding algorithms. *IEEE Transactions on Mobile Computing*, 6(6):606–620, 2007.

Kailong Chen, Tiangi Chen, Guoging Zheng, Enpeng Jin, Ou andYao, and Yong Yu. Collaborative personalized tweet recommendation. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 661–670, 2012.

Wei Chen, Laks VS Lakshmanan, and Carlos Castillo. Information and influence propagation in social networks. *Synthesis Lectures on Data Management*, 5(4): 1–177, 2013.

Justin Cheng Cheng, Lada Adamic, Alex P. Dow, Hon M. Kleinberg, and Jure Leskovec. Can cascades be predicted? In *Proceedings of the 23rd international conference on World wide web*, pages 925–936, 2014.

Radu-Ioan Ciobanu, Radu-Corneliu Marin, Ciprian Dobre, and Valentin Cristea. Interest-awareness in data dissemination for opportunistic networks. *Ad Hoc Networks*, 25:330–345, 2015.

Andrea Clementi, Riccardo Silvestri, and Luca Trevisan. Information spreading in dynamic graphs. In *Proceedings of the 2012 ACM Symposium on Principles of Distributed Computing*, PODC '12, pages 37–46, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1450-3. doi: 10.1145/2332432.2332439. URL `http://doi.acm.org/10.1145/2332432.2332439`.

Vania Conan, Jérémie Leguay, and Timur Friedman. The heterogeneity of inter-contact time distributions: its importance for routing in delay tolerant networks. 2006.

Vania Conan, Jérémie Leguay, and Timur Friedman. Characterizing pairwise inter-contact patterns in delay tolerant networks. In *Proceedings of the 1st international conference on Autonomic computing and communication systems*, page 19. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2007.

Marco Conti, Silvia Giordano, Martin May, and Andrea Passarella. From opportunistic networks to opportunistic computing. *Communications Magazine, IEEE*, 48(9):126–139, 2010.

Mario Diani and Doug McAdam. *Social movements and networks: Relational approaches to collective action*. Oxford University Press, 2003.

Savio Dimatteo, Pan Hui, Bo Han, and Victor OK Li. Cellular traffic offloading through wifi networks. In *Mobile Adhoc and Sensor Systems (MASS), 2011 IEEE 8th International Conference on*, pages 192–201. IEEE, 2011.

Sergey N Dorogovtsev, Alexander V Goltsev, and Jose Ferreira F Mendes. K-core organization of complex networks. *Physical review letters*, 96(4):040601, 2006.

Rex Yuxing Du and Wagner A Kamakura. Quantitative trendspotting. *Journal of Marketing Research*, 49(4):514–536, 2012.

Nathan Eagle and Alex Sandy Pentland. Reality mining: sensing complex social systems. *Personal and ubiquitous computing*, 10(4):255–268, 2006.

Nathan Eagle, Alex Sandy Pentland, and David Lazer. Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences*, 106(36):15274–15278, 2009.

S. B. Eisenman, E. Miluzzo, N. D. Lane, R. A. Peterson, G. S. Ahn, and A. T. Campbell. Bikenet: A mobile sensing system for cyclist experience mapping. In *ACM Transactions on Sensor Networks (TOSN), 6(1), 6*. ACM, 2009.

Károly Farkas, Theus Hossmann, Franck Legendre, Bernhard Plattner, and Sajal K Das. Link quality prediction in mesh networks. *Computer Communications*, 31 (8):1497–1512, 2008.

Willliam Feller. *An introduction to probability theory and its applications*, volume 2. John Wiley & Sons, 2008.

Emilio Ferrara and Zayao Yang. Quantifying the effect of sentiment on information diffusion in social media. In *ArXiv preprints*, 2015.

A. Foerster, K. Garg, H. A. Nguyen, and S. Giordano. On context awareness and social distance in human mobility traces. In *3rd International Workshop on Mobile Opportunistic Networks*. ACM, 2012.

Linton C Freeman. A set of measures of centrality based on betweenness. *Sociometry*, pages 35–41, 1977.

Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

Wojciech Galuba, Karl Aberer, Dipanjan Chakraborty, Zoran Despotovic, and Wolfgang Kellerer. Outtweeting the twitterers - predicting information cascades in microblogs. In *Proceedings of the 3rd Wonference on Online social networks*, pages 3–3, 2010.

Diane Gan and Lily R Jenkins. Social networking privacy–who is stalking you? *Future Internet*, 7(1):67–93, 2015.

Wei Gao and Guohong Cao. User-centric data dissemination in disruption tolerant networks. In *INFOCOM, 2011*, pages 3119–3127. IEEE, 2011.

Wei Gao, Guohong Cao, Tom La Porta, and Jiawei Han. On exploiting transient social contact patterns for data forwarding in delay-tolerant networks. *IEEE Transactions on Mobile Computing*, 12(1):151–165, 2013.

Antonios Garas, Panos Argyrakis, Céline Rozenblat, Marco Tomassini, and Shlomo Havlin. Worldwide spreading of economic crisis. *New journal of Physics*, 12(11):113043, 2010.

Antonios Garas, Frank Schweitzer, and Shlomo Havlin. A k-shell decomposition method for weighted networks. *New Journal of Physics*, 14(8):083030, 2012.

Open Gardern. Firechat application. `https://www.opengarden.com/firechat.html`, 2015.

K. Garg, S. Giordano, and A. Foerster. A study to understand the impact of node density on data dissemination time in opportunistic networks. In *HP-MoSys, 2013 Proceedings ACM*. ACM, 2013.

Kamini Garg and Silvia Giordano. Towards developing a generalized modeling framework for data dissemination. *ewsn 2015*, page 9, 2015.

Marta C Gonzalez, Cesar A Hidalgo, and Albert-Laszlo Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.

Google. Understanding consumers' local search behavior. `https://storage.googleapis.com/think-emea/docs/research_study/Report_Google_Local_Search_Behavior_DE_1.pdf`, 2015.

Robin Groenevelt, Philippe Nain, and Ger Koole. The message delay in mobile ad hoc networks. *Perform. Eval.*, 62(1-4):210–228, October 2005. ISSN 0166-5316. doi: 10.1016/j.peva.2005.07.018. URL http://dx.doi.org/10.1016/j.peva.2005.07.018.

The Guardian. Location, location, location: the rise of hyper-local marketing. https://www.theguardian.com/media-network/2015/feb/16/location-rise-hyper-local-marketing, 2015.

Haggle. The Haggle project. http://ica1www.epfl.ch/haggle/, 2006.

Bartosz Hawelka, Izabela Sitko, Pavlos Kazakopoulos, and Euro Beinat. Collective prediction of individual mobility traces with exponential weights. *arXiv preprint arXiv:1510.06582*, 2015.

JAP Heesterbeek. *Mathematical epidemiology of infectious diseases: model building, analysis and interpretation*, volume 5. John Wiley & Sons, 2000.

Tuan-Anh Hoang and Ee-Peng Lim. Virality and susceptibility in information diffusions. In *Sixth International AAAI Conference on Weblogs and Social Media*, 2012.

Liangjie Hong, Aziz Doumith, and Brian D. Davison. Personalized retweet prediction in twitter. In *4th Workshop on Information in Networks*, 2012.

T. Hossmann, F. Legendre, P. Carta, P. Gunningberg, and C. Rohner. Twitter in disaster mode: Opportunistic communication and distribution of sensor data in emergencies. In *In Proceedings of the 3rd Extreme Conference on Communication: The Amazon Expedition*. ACM, 2011a.

Theus Hossmann, Thrasyvoulos Spyropoulos, and Franck Legendre. Putting contacts into context: Mobility modeling beyond inter-contact times. In *Proceedings of the 12th ACM International Symposium on Mobile Ad Hoc Networking and Computing*, page 18. ACM, 2011b.

W-J Hsu, Thrasyvoulos Spyropoulos, Konstantinos Psounis, and Ahmed Helmy. Modeling time-variant user mobility in wireless mobile networks. In *IEEE INFOCOM 2007-26th IEEE International Conference on Computer Communications*, pages 758–766. IEEE, 2007.

Wei-Jen Hsu, Thrasyvoulos Spyropoulos, Konstantinos Psounis, and Ahmed Helmy. Modeling spatial and temporal dependencies of user mobility in wireless mobile networks. *IEEE/ACM Transactions on Networking (ToN)*, 17(5): 1564–1577, 2009.

Pan Hui, Jon Crowcroft, and Eiko Yoneki. Bubble rap: Social-based forwarding in delay tolerant networks. In *Proceedings of the 9th ACM International Symposium on Mobile Ad Hoc Networking and Computing*, MobiHoc '08, pages 241–250, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-073-9. doi: 10.1145/1374618.1374652. URL http://doi.acm.org/10.1145/1374618.1374652.

Kazem Jahanbakhsh, Valerie King, and Gholamali C Shoja. Predicting human contacts in mobile social networks using supervised learning. In *Proceedings of the Fourth Annual Workshop on Simplifying Complex Networks for Practitioners*, pages 37–42. ACM, 2012.

Márk Jelasity, Alberto Montresor, and Ozalp Babaoglu. Gossip-based aggregation in large dynamic networks. *ACM Transactions on Computer Systems (TOCS)*, 23(3):219–252, 2005.

Evan PC Jones, Lily Li, Jakub K Schmidtke, and Paul AS Ward. Practical routing in delay-tolerant networks. *IEEE Transactions on Mobile Computing*, 6(8):943–959, 2007.

Philo Juang, Hidekazu Oki, Yong Wang, Margaret Martonosi, Li Shiuan Peh, and Daniel Rubenstein. Energy-efficient computing for wildlife tracking: Design tradeoffs and early experiences with zebranet. In *ACM Sigplan Notices*, volume 37, pages 96–107. ACM, 2002.

Thomas Karagiannis, Jean-Yves Le Boudec, and Milan Vojnović. Power law and exponential decay of inter contact times between mobile devices. In *Proceedings of the 13th Annual ACM International Conference on Mobile Computing and Networking*, MobiCom '07, pages 183–194, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-681-3. doi: 10.1145/1287853.1287875. URL http://doi.acm.org/10.1145/1287853.1287875.

Patrick Gage Kelley and Justin Cranshaw. Conducting research on twitter: A call for guidelines and metrics. 2013.

Chul-Ho Lee et al. Heterogeneity in contact dynamics: helpful or harmful to forwarding algorithms in dtns? In *Modeling and Optimization in Mobile, Ad*

*Hoc, and Wireless Networks, 2009. WiOPT 2009. 7th International Symposium on*, pages 1–10. IEEE, 2009.

Jure Leskovec, Lada A Adamic, and Bernardo A Huberman. The dynamics of viral marketing. *ACM Transactions on the Web (TWEB)*, 1(1):5, 2007.

Yong Li, Mengjiong Qian, Depeng Jin, Pan Hui, Zhaocheng Wang, and Sheng Chen. Multiple mobile data offloading through disruption tolerant networks. 2013.

David Liben-Nowell and Jon Kleinberg. The link-prediction problem for social networks. *Journal of the American society for information science and technology*, 58(7):1019–1031, 2007.

Linyuan Lü, Duan-Bing Chen, and Tao Zhou. The small world yields the most effective information spreading. *New Journal of Physics*, 13(12):123005, 2011.

MACACO. The MACACO project: Mobile context-adaptive caching for content-centric networking. `http://macaco.inria.fr/`, 2012.

Travis Martin, Jake M. Hofman, Amit Sharma, Ashton Anderson, and Duncan J. Watts. Exploring limits to prediction in complex social systems. In *Proceedings of the 25th International Conference on World Wide Web*, pages 683–694, 2016.

Alessandro Mei, Giacomo Morabito, Paolo Santi, and Julinda Stefa. Social-aware stateless forwarding in pocket switched networks. In *Infocom, 2011 Proceedings Ieee*, pages 251–255. IEEE, 2011.

E. Miluzzo and G. Lane. Cenceme - injecting sensing presence into social networking applications. In *Proceedings of the 2nd European Conference on Smart Sensing and Context*, pages 1–28. Springer LNCS, 2007.

Damon Mosk-Aoyama and Devavrat Shah. Fast distributed algorithms for computing separable functions. *Information Theory, IEEE Transactions on*, 54(7): 2997–3007, 2008.

M. Mun, S. Reddy, K. Shilton, N. Yau, J. Burke, D. Estrin, M. Hansen, E. Howard, R. West, and P. Boda. Peir, the personal environmental impact report, as a platform for participatory sensing systems research. In *Proceedings of the 7th international conference on Mobile systems, applications, and services*, pages 55–68. ACM, 2009.

Seth A. Myers, Chenguang Zhu, and Jure Leskovec. Information diffusion and external influence in networks. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 33–41, 2012a.

Seth A Myers, Chenguang Zhu, and Jure Leskovec. Information diffusion and external influence in networks. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 33–41. ACM, 2012b.

Alex (Sandy) Pentland Nathan Eagle. Reality mining dataset. `http://realitycommons.media.mit.edu/realitymining.html`, 2006.

Jörg Ott, Esa Hyytia, Pasi Lassila, Tobias Vaegs, and Jussi Kangasharju. Floating content: Information sharing in urban areas. In *IEEE International Conference on Pervasive Computing and Communications (PerCom), 2011*, pages 136–146, 2011.

Michela Papandrea, Silvia Giordano, Salvatore Vanini, and Piergiorgio Cremonese. Proximity marketing solution tailored to user needs. In *World of Wireless Mobile and Multimedia Networks (WoWMoM), 2010 IEEE International Symposium on a*, pages 1–3. IEEE, 2010.

Andrea Passarella and Marco Conti. Analysis of individual pair and aggregate intercontact times in heterogeneous opportunistic networks. *Mobile Computing, IEEE Transactions on*, 12(12):2483–2495, 2013.

Romualdo Pastor-Satorras and Alessandro Vespignani. Epidemic spreading in scale-free networks. *Physical review letters*, 86(14):3200, 2001.

Sen Pei, Lev Muchnik, José S Andrade Jr, Zhiming Zheng, and Hernán A Makse. Searching for superspreaders of information in real-world social media. *Scientific reports*, 4, 2014.

Luciana Pelusi, Andrea Passarella, and Marco Conti. Opportunistic networking: data forwarding in disconnected mobile ad hoc networks. *IEEE Communications Magazine*, 44(11):134–141, 2006.

Yuval Peres, Alistair Sinclair, Perla Sousi, and Alexandre Stauffer. Mobile geometric graphs: Detection, coverage and percolation. In *Proceedings of the Twenty-second Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '11, pages 412–428. SIAM, 2011. URL `http://dl.acm.org/citation.cfm?id=2133036.2133069`.

Sasa Petrovic, Miles Osborne, and Victor Lavrenko. Rt to win! predicting message propagation in twitter. In *ICWSM*, 2011.

Alberto Pettarin, Andrea Pietracaprina, Geppino Pucci, and Eli Upfal. Tight bounds on information dissemination in sparse mobile networks. In *Proceedings of the 30th Annual ACM SIGACT-SIGOPS Symposium on Principles of Distributed Computing*, PODC '11, pages 355–362, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0719-2. doi: 10.1145/1993806.1993882. URL http://doi.acm.org/10.1145/1993806.1993882.

Fabio Pezzoni, Jisun An, Andrea Passarella, Jon Crowcroft, and Marco Conti. Why do i retweet it? an information propagation model for microblogs. In *Proceedings of the 5th International Conference on Social Informatics*, pages 360–369, 2013.

A. Picu, Thrasyvoulos Spyropoulos, and T. Hossmann. An analysis of the information spreading delay in heterogeneous mobility dtns. In *World of Wireless, Mobile and Multimedia Networks (WoWMoM), 2012 IEEE International Symposium on a*, pages 1–10, June 2012. doi: 10.1109/WoWMoM.2012.6263682.

Andreea Picu and Thrasyvoulos Spyropoulos. Minimum expected*-cast time in dtns. In *Bioinspired Models of Network, Information, and Computing Systems*, pages 103–116. Springer, 2010.

Qualcomm. Lte direct proximity services. https://www.qualcomm.com/invention/technologies/lte/direct, 2014.

RollerNet. The RollerNet: Analysis and use of mobility in rollerblade tours. http://rollernet.lip6.fr/en/results/results.html, 2006.

Mattew J. Salganik, Peter S. Dodds, and Duncan J. Watts. Experimental study of inequality and unpredictability in an artificial cultural market. *Science*, 311 (5762):854–856, 2006.

SCAMPI. The SCAMPI project: Service platform for social aware mobile and pervasive computing. http://www.ict-scampi.eu/results/scampi-bcards/, 2012.

Salvatore Scellato, Anastasios Noulas, and Cecilia Mascolo. Exploiting place features in link prediction on location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1046–1054. ACM, 2011.

FW Scholz. Maximum likelihood estimation. *Encyclopedia of statistical sciences*, 1985.

J. Scott, J. Crowcroft, P. Hui, and C. Diot. Haggle: A networking architecture designed around mobile users. In *Third Annual Conference on Wireless On-demand Network Systems and Services*, pages 78–86. IEEE, 2006a.

James Scott, Jon Crowcroft, Pan Hui, and Christophe Diot. Haggle: A networking architecture designed around mobile users. In *WONS 2006: Third Annual Conference on Wireless On-demand Network Systems and Services*, pages 78–86, 2006b.

James Scott, Richard Gass, Jon Crowcroft, Pan Hui, Christophe Diot, and Augustin Chaintreau. CRAWDAD data set cambridge/haggle (v. 2006-01-31). Downloaded from `http://crawdad.org/cambridge/haggle/`, January 2006c.

James Scott, Richard Gass, Jon Crowcroft, Pan Hui, Christophe Diot, and Augustin Chaintreau. CRAWDAD dataset cambridge/haggle (v. 2009-05-29). Downloaded from `http://crawdad.org/cambridge/haggle/20090529`, May 2009.

Devavrat Shah. *Gossip algorithms*. Now Publishers Inc, 2009.

Hamed Soroush, Nilanjan Banerjee, Mark Corner, Brian Levine, and Brian Lynn. A retrospective look at the umass dome mobile testbed. *ACM SIGMOBILE Mobile Computing and Communications Review*, 15(4):2–15, 2012.

Johan AK Suykens and Joos Vandewalle. Least squares support vector machine classifiers. *Neural processing letters*, 9(3):293–300, 1999.

Pierre Ugo Tournoux, Jérémie Leguay, Farid Benbadis, Vania Conan, Marcelo Dias de Amorim, and John Whitbeck. The accordion phenomenon: Analysis, characterization, and impact on dtn routing. In *Proc. IEEE INFOCOM*, 2009.

Salvatore Vanini, Claudio Di Giacinto, Dario Gallucci, Tiziano Leidi, Silvia Giordano, and Piergiorgio Cremonese. User-profiled platform with advanced navigation support. In *World of Wireless, Mobile and Multimedia Networks (WoWMoM), 2012 IEEE International Symposium on a*, pages 1–3. IEEE, 2012.

Gary Vaynerchuk. How to master the 4 big social-media platforms. `http://www.inc.com/magazine/201311/gary-vaynerchuk/how-to-master-the-four-major-social-media-platforms.html`, 2013.

Vladimir Vukadinovic and Stefan Mangold. Opportunistic wireless communication in theme parks: a study of visitors mobility. In *Proceedings of the 6th ACM workshop on Challenged networks*, pages 3–8. ACM, 2011.

Duncan J Watts, Jonah Peretti, and Michael Frumin. *Viral marketing for the real world*. Harvard Business School Pub., 2007.

Sanford Weisberg. *Applied linear regression*, volume 528. John Wiley & Sons, 2005.

John Whitbeck, Marcelo Amorim, Yoann Lopez, Jeremie Leguay, and Vania Conan. Relieving the wireless infrastructure: When opportunistic networks meet guaranteed delays. In *World of Wireless, Mobile and Multimedia Networks (WoWMoM), 2011 IEEE International Symposium on a*, pages 1–10. IEEE, 2011.

Jiang Yang and Scott Counts. Predicting the speed, scale, and range of information diffusion in twitter. In *In 4th International AAAI Conference on Weblogs and Social Media*, 2010.

Eiko Yoneki. Fluphone study: Virtual disease spread using haggle. In *Proceedings of the 6th ACM Workshop on Challenged Networks*, CHANTS '11, pages 65–66, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0870-0. doi: 10.1145/2030652.2030672. URL http://doi.acm.org/10.1145/2030652.2030672.

Nicholas J. Yuan, Yuan Zhong, Fuzheng Zhang, Xing Xie, Chin-Yew Lin, and Yong Rui. Who will reply to/retweet this tweet?: The dynamics of intimacy from online social interactions. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, pages 3–12, 2016.

Jian-Xiong Zhang, Duan-Bing Chen, Qiang Dong, and Zhi-Dan Zhao. Identifying a set of influential spreaders in complex networks. *arXiv preprint arXiv:1602.00070*, 2016.

Xiaolan Zhang, Giovanni Neglia, Jim Kurose, and Don Towsley. Performance modeling of epidemic routing. *Comput. Netw.*, 51(10):2867–2891, July 2007. ISSN 1389-1286. doi: 10.1016/j.comnet.2006.11.028. URL http://dx.doi.org/10.1016/j.comnet.2006.11.028.

Z. Zhang. Routing in intermittently connected mobile ad hoc networks and delay tolerant networks: overview and challenges. In *Communications Surveys and Tutorials, IEEE*, pages 24–37. IEEE, 2006.

Huan Zhou, Jiming Chen, Jialu Fan, Yuan Du, and Sajal K Das. Consub: incentive-based content subscribing in selfish opportunistic mobile networks. *IEEE Journal on Selected Areas in Communications*, 31(9):669–679, 2013.

Xia Zhou, Stratis Ioannidis, and Laurent Massoulié. On the stability and optimality of universal swarms. *ACM SIGMETRICS Performance Evaluation Review*, 39 (1):301–312, 2011.