UNIVERSITY OF CALIFORNIA
RIVERSIDE

Measuring and Inferring Demographics from Multi-modal Communication Networks

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Computer Science

by

Yi Wang

August 2013

Dissertation Committee:

Professor Michalis Faloutsos, Chairperson
Professor Eamonn Keogh
Professor Stefano Lonardi
Professor Vassilis Tsotras

The Dissertation of Yi Wang is approved:

Committee Chairperson

University of California, Riverside

Acknowledgments

I am grateful to my supervisor, Prof. Michalis Faloutsos, whose expertise, understanding, generous guidance and support made it possible for me to work on a topic that was of great interest to me. It was a pleasure working with him.

I am hugely indebted to Dr. Hui Zang for giving her precious and kind advice regarding the topic of my research and Sprint for providing me with the datasets that I could not possibly collect on my own.

I would like to express my gratitude to Dr. Marios Iliofotou for his immense interest in my topic of research and for finding time for us in his busy schedule. Sir, words can never be enough to thank your kindness.

I would like to express my gratitude to my dissertation committee members: Prof. Eamonn Keogh, Prof. Stefano Lonardi and Prof. Vassilis Tsotras for their advices on the dissertation.

I am particularly thankful for the support of my family and friends, for everything that has led me to this point.

ABSTRACT OF THE DISSERTATION

Measuring and Inferring Demographics
from Multi-modal Communication Networks

by

Yi Wang

Doctor of Philosophy, Graduate Program in Computer Science
University of California, Riverside, August 2013
Professor Michalis Faloutsos, Chairperson

With the proliferation of electronic modes of communication (e.g., e-mails, phone calls and short messages), a group of people can form several distinct communication networks. A communication network is essentially a graph representation of who talks (or texts) to whom among a group of individuals. In this thesis, we conduct an empirical study of communication networks in two modern countries and focus on four questions: 1) what are the patterns of multimodal communication across countries and over time? 2) how much information can we extract regarding the roles of users? 3) can we infer the demographic properties of certain users? 4) can we predict the phone choices for a group of users? For the first question, I study the correlation between calling and texting across two countries: China and the U.S., and the evolution of the usage of the two communications over the last five years. I propose to depend on communication channels (calling and texting) and time-slices (weekday, weekend, and holiday) to study how people in China and the U.S. contact one another. This idea is inspired by the fact that different human relationships

can be indicated by communications in different time slicess. For example, texting tends to indicate a friendship, while calling on a weekday morning could indicate a relationship between colleagues. For the second question, I examine the effect of communication channels on role prediction. I first show the similarity and differences in calling and texting between managers and ordinary employees and then propose a ranking algorithm, called HumanRank, which infers employee role of the job title with 10% higher accuracy than existing methods. I discuss how the texting graph is 10% better at role prediction than the calling graph. For the third question, I explore the effect of time slices on the prediction of age group and income level on call networks by studying the correlation between calling features and the demographic homophily and then proposing a prediction algorithm to reach an accuracy of 80% for age group and 71% for income level. Moreover, I discuss how the weekday graph is 15% better than the night-weekend graph regarding the prediction accuracy. for the fourth question, I study correlations between demographics and phone preference. With the features emerging from the correlations, I devise a solution to infer a user's phone choice. Compared with existing methods, my solution reduces the error by 1/3 and related costs by half.

Contents

List of Figures **ix**

List of Tables **xii**

1 Introduction **1**
1.1 Overview 5

2 The patterns of multimodal communication across countries and over time **6**
2.1 Introduction 6
2.2 Framework and definitions 9
2.3 Related work 12
2.4 Statistics and Correlations 12
2.4.1 Are calling and texting correlated? 13
2.5 Cross-cultural study 16
2.5.1 More calls than texts in 2007 16
2.5.2 Periodicity and intra-day peaks 17
2.5.3 Holidays is the season for texting 18
2.5.4 China adopts texting earlier than the U.S. 19
2.6 Time evolution study 21
2.7 Conclusions 23

3 Extraction of information regarding the roles of users **25**
3.1 Introduction 25
3.2 Related Work 28
3.3 Datasets and Graph Metrics 30
3.3.1 Datasets 30
3.3.2 Metrics for analyzing and comparing graphs 31
3.3.3 Generating synthetic random networks for comparison 33
3.4 Characterizing Enterprise CINs 35
3.4.1 Comparing the CALL, SMS, and EMAIL CINs 35

3.4.2 Comparing CINs with popular complex networks 39
3.4.3 Comparison with synthetic networks 40
3.5 Communication patterns in CALL and SMS 44
3.6 Detecting Hierarchy Structure in Enterprise CINs 47
3.6.1 HumanRank: Ranking employees in enterprises using CINs 47
3.6.2 Hierarchy detection using HumanRank 50
3.6.3 Experimental Evaluation 51
3.7 Summary and Conclusions 54

4 Inference of the demographic properties of users **56**
4.1 Introduction 56
4.2 Problem definition and data introduction 59
4.3 Observations 62
4.3.1 Quantifying graph homophily 62
4.3.2 Communication features as indicators of homophily 65
4.4 Inference of home location, age and income 68
4.5 Conclusions 72

5 Patterns of Phone Switching and Inference **73**
5.1 Introduction 73
5.2 Related Work 75
5.3 Datasets and features 77
5.3.1 Dataset introduction 77
5.3.2 Definition of features 79
5.4 Observations and trends 80
5.5 Predicting phone switches 86
5.6 Summary and Conclusions 92

6 Conclusion **94**

Bibliography **97**

List of Figures

2.1 The joint probability of (a) Number of texts and calls (b) Text degree and call degree 13
2.2 Texting and calling temporal patterns: number of calls and texts per hour over a week of (a) December 2007 in a Chinese city (b) December 2007 in Greater San Francisco Area. In (a), the numbers of calls peak at 10 a.m. and at 5 p.m. and that of texts is at 4 p.m.. 17
2.3 Texting and calling temporal patterns: number of calls and texts per hour over a week of (c) July 2011 in Greater San Francisco Area. In 2007, the numbers of calls peak at 10 a.m. and at 5 p.m. and that of texts is at 4 p.m.. In 2011, calls burst at 6 p.m.. 18
2.4 Number of records per day with the holidays indicated. On Christmas Day and New Year's Day, the number of texts increases by 100% in the Chinese city 20
2.5 The evolution of (a) communication events, and (b) users for users active in that mode of communication over 4 years in the greater San Francisco area. The number of active texters per day surpasses that of callers in 2009. D1 = 12/08/07, D2 = 08/26/08, D3 = 03/11/09, D4 = 07/07/09, D5 = 04/14/10, D6 = 02/28/11, D7 = 07/16/11 22
2.6 The evolution of average daily activity for users active in that mode of communication over 4 years in the greater San Francisco area. The number of active texters per day surpasses that of callers in 2009. D1 = 12/08/07, D2 = 08/26/08, D3 = 03/11/09, D4 = 07/07/09, D5 = 04/14/10, D6 = 02/28/11, D7 = 07/16/11 24

3.1 Comparison of the CALL, SMS, and EMAIL CINs using six different no-scalar metrics. All networks have similarities in their degree distributions, clustering coefficient, path lengths, and their degree-to-degree correlations. The SMS and EMAIL are more similar to each than to the CALL graph. . 36

3.2 Comparing the CALL network with four synthetic graphs. The CALL CIN is different from random graphs, especially in the clustering and the formation of large cliques. From all the synthetic graphs, the power-law (BA) is the most different from our CALL CIN. 42
3.3 In the scatter plots we compare the characteristics of each employee in the communication provider's enterprise using the CALL (x-axis) and SMS (y-axis) CINs. In (a) and (b) we compare the degrees and clustering coefficients for the managers (crosses) and ordinary employees (dots), for the CALL and SMS networks respectively. In (c) we compare the number of phone calls made and short messages send by each employee over the duration of our dataset. 45
3.4 A Venn diagram comparing the number of communicating pair of nodes that use: (a) only short messages, (b) only phone calls, and (c) both short messages and phone calls. From the diagram we see than most of the communicating node pairs use a single-mode communication and only 41% are observed to use multi-mode communications. 47
3.5 The two step process for labeling nodes as ordinary employees or managers. All the nodes are firstly ranked according to HumanRank and then are assigned in two groups using k-means (k=2). 51
3.6 Ranking results: The plots show the number of managers identified in the top k nodes, as ranked by our HumanRank and three other ranking methods. We also compare our results with an optimal ranker where all managers have higher score than ordinary employees. PageRank is not applied on SMS because SMS is undirected. 52
3.7 Accuracy and Manager's F-score in hierarchy detection using HumanRank and K-means (k=2). We also show measure results of two supervised learning methods, and one unsupervised method using the ranking from [1] and K-means (k=2). 53

4.1 The probabilities $p_1(d)$ (LEFT) and $p_2(d)$ (MIDDLE) as a function of d and Cumulative Distribution Function (CDF) of $p_2(d)$ (RIGHT). 63
4.2 People talk to people of same age group and income level with more than 50% probability: Probability of communication among (LEFT) age groups, and (RIGHT) income levels. The x-axis presents a and y-axis presents b. . 63
4.3 Complementary cumulative distribution functions (CCDF) of (LEFT) the percentage, and (RIGHT) the absolute number of neighbors of the same age/income. 66

5.1 The probability that a subscriber uses a phone PH_{curr} given (top) his/her age group and (bottom) income level. 82
5.2 Probability Distribution Function of the length Len of the life of a used phone PH_{prev} over (top) all subscribers and (bottom) rich and poor subscribers. . 83
5.3 The probability that a subscriber uses the same phone PH_{curr} as the used one PH_{prev} given the length of the life of the used phone. 84

5.4 Probabilities of (top) PH_{social}=PH_{curr} and (bottom) PH_{plan}=PH_{curr} for the size of a neighborhood from 2 to 20. 87
5.5 The absolute difference (AD) and the absolute cost different (ACD) between groundtruth and the prediction by Histro, Baye, Baye+Mat, and Logi. . . . 91

List of Tables

2.1 Basic statistics of dataset 10
2.2 Correlation across channels. C = CALL, T = TEXT, WD = WEEKDAY, HD = HOLIDAY, WE =WEEKEND. Two users who have texting on weekends are more likely (more than 50%) to communicate with each other at other channels. 16

3.1 Scalar graph metrics for CALL, SMS and EMAIL CINs compared to real and synthetic Networks. $|V|$: number of nodes, ED: Edge density, $\bar{k}$: average degree, D: Diameter. P: Path length and $\bar{C}$: average clustering coefficient . 38
3.2 Presentation of selected directed motifs of size three. 43

4.1 Basic statistics of data sets 60
4.2 The results of inferring demographic properties 69
4.3 Confusion matrix of inferences IDs = 9 and 15. ACT=Actual, PRE=Predicted 71

5.1 An example of the phone-switching history of a subscriber 78
5.2 Definitions of symbols 81
5.3 Probability of switching from the used phone PH_{prev} to the current phone PH_{curr} . The first row and column represent PH_{curr} and PH_{prev}. 85
5.4 Result of predictions by Histro, Logistic regression and Bayesian network with/without cost matrix. $\downarrow$ and $\uparrow$ mean under- and over-estimating. . . . 90
5.5 Under- and over-estimated costs of 6 types of phones 90

Chapter 1

Introduction

Today‘s proliferation of electronic modes of communication (e.g., phone calls, text messages, and e-mails) are changing the way people read, learn, and interact. More and more people are engaging in electronic activities for an unprecedented level of convenience. The number of mobile phone users in the U.S. has increased from 34 million to 203 million in the last ten years. Every year, there are 2.12 trillion text messages sent. Smartphones have also shown a significant growth worldwide over the past few years. Estimates say that there are now more than 1.5 billion smartphones in the world, and about 81% of mobile users are using 3G/4G services via smartphones in the U.S..

The explosion in the use of electronic communications also record the ways in which people interact with each other and store them in form of system logs. By analyzing these logs, researchers have been able to study and answer a fundamental question of “how people communicate with others, where, when, and in what kind of way”, from various aspects across diverse areas of computer science. For example, in the area of data mining,

some researchers are exploring the answer utilizing graphs. In the area of networking, researchers are finding the solution from studying mobility patterns.

In this thesis, I use Call Detail Records (CDR), where each line records information of a call event, including its caller and callee, both parties' geographical locations, and the time of the event. I explore the answer to the general question and study it by inferring personal attributes. Generally speaking, the fundamental question of the thesis is "what attributes of people can we predict from the CDR?" I answer this general question from four aspects: multimodal communication, detection of user role, inference of demographics, and prediction of phone choices. Each of them corresponds, respectively, to one question:

1. What are the patterns of multimodal communication across countries and over time?
2. How much information can we extract regarding the roles of users?
3. Can we infer the demographic properties of certain users?
4. Can we predict the phone choices for a group of users?

For the first question, I present correlations in multimodal communications, the effect of culture on communication patterns, and the evolution of multimodal communications over the last five years. The goal of this question is to show the basic and aggregated results from citywide CDR datasets so that people can obtain straightforward knowledge about the "shape" of multimodal communications. For the second question, I show how people in the same company contact one another via multimodal communications and how to utilize their communication patterns to infer an employee's job title. For the third question, I am interested in the ability to infer age group, income level, and home locations of a

user and discuss how to depend on a user's "best friends" to infer his/her demographics. For the last question, I will show correlations between user demographics and phone preference and present a solution to the problem of predicting the phone a subscriber will use next.

The contributions of my dissertation to existing research are as follows. First, I study the correlation between calling and texting across two countries: China and the U.S., and the evolution of the usage of the two communications over the last five years. I propose to depend on communication channels (calling and texting) and time-slices (weekday, weekend, and holiday) to study how people in China and the U.S. contact one another. This idea is inspired by the fact that different human relationships can be indicated by communications in different time slices and channels and utilizing this separation will increase the accuracy of prediction algorithms. For example, texting tends to indicate a friendship, while calling on a weekday morning shows a relationship between colleagues.

Second, I examine the effect of communication channels on title prediction. I first show the similarity and differences in calling and texting between managers and ordinary employees and then propose a ranking algorithm, called HumanRank, which infers job title with 10% higher accuracy than existing methods. I discuss how the texting graph is 10% better at title prediction than the calling graph.

Third, I explore the effect of time slices on the prediction of age group and income level on call networks by studying the correlation between calling features and the demographic homophily and then proposing a prediction algorithm to reach an accuracy of 80% for age group and 71% for income level. Moreover, I discuss how the weekday graph is 15% better than the night-weekend graph regarding the prediction accuracy.

Fourth, I study correlations between demographics and phone preference. With the features emerging from the correlations, I devise a solution to infer a user's phone choice. Compared with existing methods, my solution reduces the error by 1/3 and related costs by half.

My dissertation belongs to an interdisciplinary work between two areas: networking and data mining. On the one side, I utilize classical data mining methods to solve a series of new problems appearing in networking. More concretely, my work heavily shows a selection of features that are specific to the domain of networking and discusses how to use those features through classical data mining methods to deal with a problem. On the other side, my work extends and adopts powerful data mining techniques and show their ability to infer and predict useful properties of human behavior.

Limitation. My work focuses specially on the communications through calling and texting. I do not consider communications via Online Social Networks, like Facebook and Twitter. The focus of my work is reflected by the following aspects. First, we use only CDR and its associated datasets to answer all the questions. Our CDRs are from two countries: China and the U.S. Second, the underlying social network extracted from the CDR is actually a call graph in which one node is a caller or a callee and an edge represents a call event. The call graph is not able to capture the completeness of human interactions. Note that the definition of the graph will be different in each chapter, so please refer to each chapter for detailed definitions.

1.1 Overview

This dissertation is an effort to serve as an introduction to the measurement of multi-modal communication networks as well as inference of demographic properties of network subscribers, including home location, company title, age group, income level, and phone preference. The general organization of this dissertation is as follows. Chapter 2 provides the patterns of multimodal communication across countries and over time. Chapter 3 discusses extraction of information regarding the roles of users from an enterprise communication network. Chapter 2 and 3 use both calling and texting datasets. Chapter 4 discusses inference of the users' demographic properties. Chapter 5 studies patterns of phone switching and algorithms for predictions of phone preferences. Chapter 4 and 5 use only the calling datasets. Chapter 6 provides a summary of the work.

Each chapter is organized in the measurement-inference style with 2-3 sub-questions in the interest of answering the general questions. In other words, at first, I discuss various observations and then, based on these observations, I discuss how to design the inference algorithms. For example, in answering the second question, I will show the similarities and differences of how multimodal communications are used among ordinary employees and managers and then discuss how extensively one should rely on the differences to infer their company titles.

Chapter 2

The patterns of multimodal communication across countries and over time

2.1 Introduction

So far, most research efforts have studied different communication modes in *isolation*. Wireline calling patterns in isolation have been analyzed extensively over the last 100 year and recently mobile calling patterns[2]. Relatively fewer studies have analyzed texting[3], again in isolation. Other communication modes, such as emailing, and instant messaging, have been study but these are not modes we study here. To the best of our knowledge, there has not been a study focusing on the composite communication network of calls and texts, as we do here. In section 2.3, we provide an overview of previous work.

Modern communications between human beings are multimodal, so looking at only one mode is providing an incomplete picture of the communication pattern. Here, we focus on text and call multimodal networks and ask three key questions:

a. How do we capture multimodal correlations?

b. How does culture affect communication patterns? chapter *c. How has it evolved over the last five years?*

An underlying challenge, multimodal analysis for human-centric networks requires some new descriptive terms and metrics to quantify interesting behaviors, as we elaborate below.

The contributions of this chapter are threefold. First, we present some initial ideas towards a framework for analyzing multimodal communication networks, which we refer to as **MCN**s. Second, we study the effect of culture on multimodal communications. Third, we observe the evolution of texts and calls within 5 years. We use three large datasets in our study obtained from cellphone service providers: (a) a Chinese dataset (CN-07) in 2007-2008, with 154 thousand users and 3 million records, from the captial of a province in China, (b) the great San Francisco dataset (SF-07) in 2007-2008 with 1 million users and 450 million records, and (c) the San Francisco dataset (SF-11) in 2011 with 1.4 million users and more than 1 billion records.

As a first step towards a systematic study, we find it necessary to introduce a set of definitions and metrics to describe and capture interesting aspects of the communication duality. First, we suggest that it is important to study behavior at both user and communicating pair levels, namely two users that engage in a communication. Second, we introduce

the concepts of **time-slices**, essentially time-based windows, and **channels**, each of which captures a mode of communication over a specific time-slice. For example, we propose three time-slices: weekday, weekend and holiday, and we define six channels to be pairs of (mode, time-slice) e.g. (text, weekend). Third, we propose techniques to assess the correlation between an activity in one channel with the existence of the activity in another channel.

Our key observations can be summarized as follows.

A. Dual-mode statistics and Correlations: Our goal is to understand and quantify the different ways people use the different communication modes.

a. 51% of the people in CN-07 and 48% in SF-07 use both texts and calls.

b. Aggregate user behavior is stable along three time-slices: weekday, weekend, and holiday. The key observation is that the behavior is (a) repeatable and stable, and (b) different from each other. This suggests that studying the behavior along those time-slices is meaningful and potentially useful at level of granularity.

c. For a communicating pair, we find that the existence of texting during the weekend time-slice is the strongest indicator of persistent communication in terms of the likelihood that the pair interacts in other channels.

B. Cultural comparison: Our goal is to quantify whether different cultures or countries exhibit significantly different multimodal communication patterns.

a. Cultural differences: Chinese users exhibit dual peaks for calls (at 10 a.m. and 5 p.m.) and a single peak for text (at 4 p.m.) on weekdays. In contrast, the U.S. users have a single peak for calls (at 6 p.m.) and no pronounced peak for texts. Holidays exhibit spikes in texts for both countries, but by different levels of 100% increase in China, and

30% in the U.S.. Surprisingly, the number of calls decreases during one major holiday in China. Moreover, Chinese users text more than the U.S. users in terms of the percentage of texters (68% v.s. 52%).

b. Cultural similarities: The number of callers is 2.7 times that of texters on a single day for both countries in 2007. We have also seen obvious daily and weekly periodicities, and a similar decrease of activities during weekends for both countries.

C. Evolution: Texting has started to dominate. We find that texting has taken over in sheer number of events by flipping the ratio of number of calls over number of texts ratio from 2:1 in 2007 to 1:2 in 2011 based on the U.S. datasets. Investigating this further, we find an increase in both the number of texts per texting user, and the number of people texting: the percentage of texters is up to 74% in 2011 from 52% in 2007 of all users. Interestingly, although the number of texters has increased, it is still less than the number of callers: from 1:3 in 2007 to 2:3 in 2011.

2.2 Framework and definitions

We model communications as an undirected network, where nodes are cellphone numbers and links correspond to one or more communication events, which could be either a text or a call. A **record** or **communication event** relates two participating users and the time of occurrence. We use the term **user** or **individual** to refer to a cellphone number, although one phone can be shared among different people. We use the terms **caller** or **texter** to refer to users making at least one call or text. Note that a caller can also be a texter, as we discuss below.

Table 2.1: Basic statistics of dataset

	Time range	#calls	#texts	#users	#callers	#texters
CN-07	12/07-03/08	2,065K	1,208K	154K	128K	105K
SF-07	12/07-03/08	273,245K	189,983K	1,065K	1,043K	532K
SF-11	06/11-08/11	327,136K	629,045K	1,453K	1,301K	1,097K

Framework, definitions and concepts. We introduce a set of definitions and metrics in order to analyze Multimodal Communication Network effectively.

First, we find important to study the behavior at both the user level and the communicating pair level, essentially an edge of the network. For example, a **dual-mode user** is a user who is both a caller and a texter, while a **dual-mode communicating pair** uses both modes to communicate with each other.

Second, we introduce the concepts of **time-slices** and **channels**, as we mentioned in the introduction. which capture modes of communication over distinct time-based windows in a human centric way that we call time-slice s. We argue that it useful to have three time-slices: weekday, weekend and holiday, since we see different behaviors over those time-frames. We can further define channels to be pairs of (mode, time-slice) e.g. (text, weekend). We show that their correlations and the nature of the communication differs across the different channels.

Finally, we propose different metrics and techniques to assess the correlation between an activity in one mode or channel with the existence of the activity in another channel. Due to space limitations, we can only show a small subset here.

Datasets of this study: We use three large datasets in our study obtained from cellphone service providers: (a) CN-07: a Chinese dataset between Dec-2007 and March-2008, with 154 thousand users and 3 million records, (b) SF-07: the greater San Francisco area between Dec-2007 and March-2008 in 2007-2008 with 1 million users and 450 million records, (c) SF-11: the greater San Francisco area data in 2011 with 1.4 million users and more than 1 billion records, and (d) SF-days: San Francisco data on several distinct days: August 26th 2008, March 11th 2009, July 7th 2009, April 14th 2010 and Feburary 28th 2011. For studying cultural differences, we use the CN-07 and SF-07 datasets, which include Christmas Day and New Year and the Chinese New Year. For studying the evolution, we use the SF-07, SF-11, and SF-days datasets.

User selection filter. We prefer to reduce the number of users we analyze to ensure homogeneity, and avoid introducing statistical artifacts. First, our data only consists of cellphone users, so that any two nodes can call and text[1]. Second, we distinguish between two types of users in our dataset: (a) customers of the telecom carrier, which provide user logs, and (b) non-customers who communicate with customers. Here, *we only consider customers of two telecom carriers as users.* The reason is that in profiling a user we want to be sure that we see all the activity of each user, which is only true for the customer belonging to the two carriers. Studying how customers of one carrier interact with customers of other carriers could be an interesting study in its own right.

[1]We assume that all cellphones are text capable, which is for all practical purposes a reasonable assumption.

2.3 Related work

Many research efforts have studied communication modes in isolation: phone calls[4], text messages[3], email and instant messaging[5]. Most previous works [6], [2], [3] on cellphone networks, study calls and texts in isolation, while we study the interplay of these two modes. Some studies focus on topology features of a larger-scale phone-call network[6], while other studies focus on tie persistence in a text message networks in order to detect anomalies [3, 7]. Characterizing mobile phone networks can help telecom carriers to better manage and provision their networks[8][9], and we believe that studying the communication patterns jointly could help this further, and potentially also inspire user friendly call and text end-user packages.

2.4 Statistics and Correlations

In this section, we analyze the CN-07 dataset on both user and pair levels. We study the evolution of texts and calls in Section 2.6.

Both text and calls exhibit three stable time-slices. We observe three distinct time-slices: weekday (**WD**), weekend (**WE**) , and holiday (**HD**). A sample of this stable behavior is shown in Figure **??** in section 2.5 where we discuss its properties in more detail. Due to space limitations, we cannot show more instances, but this behavior was very consistent. Apart from the weekly periodicity, daily activity was very predictable in terms of spikes. Furthermore, weekends always exhibit lower activity than weekdays. This suggests that studying the behavior along those time-slices is meaningful and provides a

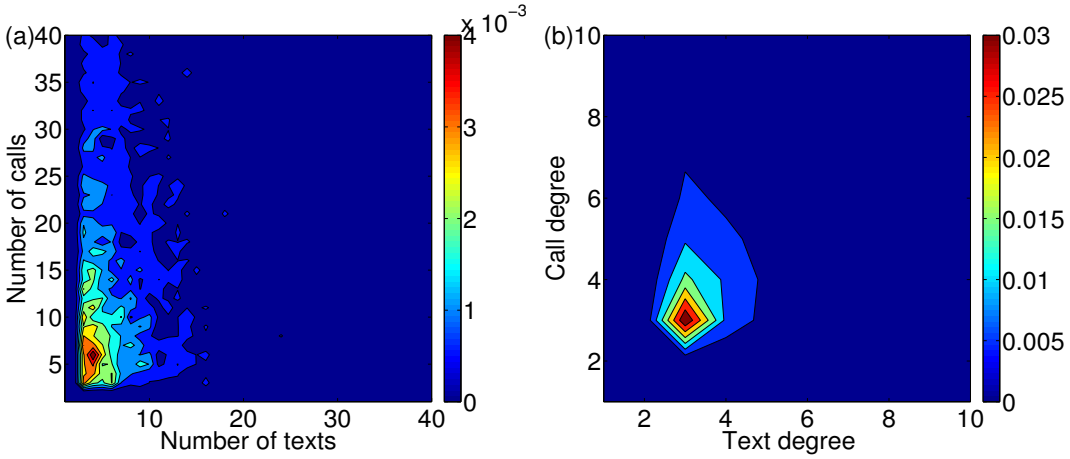


Figure 2.1: The joint probability of (a) Number of texts and calls (b) Text degree and call degree

useful level of granularity. Given these three time-slices, we can define six channels, by associating a communication mode with a time-slice: for example, (text, WD), (call, HD) etc.

Dual-mode behavior is fairly limited: Only half of the people (51%) in the CN-07 dataset are dual-mode users. Specifically, the user base includes 51% dual-mode users, 32% use calls exclusively, and 17% use texts exclusively. Going a step further, we analyze communication pairs and we find that only 23.7% are dual-mode pairs, while slightly more than half of the pairs (58.7%) use exclusively calls, and 17.6% use texts exclusively.

2.4.1 Are calling and texting correlated?

Here, the goal is to quantify the correlation between the two modes of communication. We start by looking at two level analysis: (a) number of events, and (b) number

of neighbors that each mode is used on. First, we study the numbers of calls and texts of dual-mode users (both caller and texter) and plot the joint probability shown in Figure 2.1 (a). The correlation coefficient is 0.31, indicating a weak positive correlation. We see that $> 85\%$ of the users make more calls than texts, as indicated by the larger lighter area above diagonal than below it.

We identify some less common, but quite surprising behaviors. First, we find a group of users with very low number of calls (<10) but very high number of texts (>500). Some of them send texts to 400 persons per day, which very likely text spamming, especially given the really low calling activity. Note that, for visual clarity, these users are beyond the visible scale in figure which focuses on the user majority. Although this extends beyond the scope of this chapter, this could be a good first step in identifying suspicious behavior using correlation information.

In Figure 2.1 (b), we study the number of partners that a dual-mode user calls (or call degree) and the number of partners who the same user texts (text degree). The joint probability of the call degree and text degree is shown in Figure 2.1 (b), and exhibits weak correlation with a 0.30 coefficient. It is fairly easy to see that most of the users are near the diagonal, compared to the number of calls and texts that a user engages in.

Inferring texting from calling and vice versa. We attempt to delve deeper into the correlations of the texting and calling behaviors for two questions: (1) *if I call you, do I text you?* (2) *If I text you, do I call you?* Namely, we investigate the predictive power of one mode of communication for the existence of the other mode in a communicating pair.

We formalize this by employing the conditional probability of the appearance of

an edge, r, in a network (N_2) given that it exists in network (N_1) based on the observed behavior. We consider that a communicating pair can be connected with two types of edges texting edge, a calling edge, or both edges. We calculate this as shown below, with $edges()$ indicating the edges of a network:

$$P(r \in N_2 | r \in N_1) = \frac{|edges(N_1) \bigcap edges(N_2)|}{|edges(N_1)|}$$

In the above equation, we can consider N_1 and N_2 to represent the call network and the text network. We find that the probability P(CALL|TEXT) is 57.3% while P(TEXT|CALL) is 28.7%. This shows that a pair that texts is likely to also call, but not the other way around.

Correlation across time-slices and channels. We go one step further and analyze the conditional probabilities between behaviors over different time-slices (weekend, weekday and holiday). In Table 2.2 we show all the conditional probabilities. The way to read this table is to start from the row, say TWE, and match it with a column, say CWD: *if I text you on weekends, do I call you on weekdays?*. The answer is with 0.63 probability.

The first observation is that texting on the weekend (TWD) seems to be the strongest indicator that there will be communication in all other channels: the row TWD shows high values, with the only exception that the correlation between TWD and calling on holidays (CHD) is as low as 0.29. In other words, TWD implies strong multimodal relationships in the other channels.

The second observation is that almost all communication channels are good indicators that the pair will also call on weekdays (CWD): the column CWD exhibits high values. In other words, if a pair is communicating in any other channel, the chance is high

Table 2.2: Correlation across channels. C = CALL, T = TEXT, WD = WEEKDAY, HD = HOLIDAY, WE =WEEKEND. Two users who have texting on weekends are more likely (more than 50%) to communicate with each other at other channels.

	CWD	CWE	CHD	TWD	TWE	THD
CWD	•	0.33	0.18	0.17	0.08	0.19
CWE	0.68	•	0.28	0.22	0.15	0.24
CHD	0.72	0.54	•	0.22	0.15	0.32
TWD	0.63	0.39	0.23	•	0.31	0.43
TWE	**0.63**	**0.55**	**0.29**	**0.64**	•	**0.50**
THD	0.45	0.28	0.19	0.28	0.15	•

that it will also be calling on weekdays.

2.5 Cross-cultural study

In this section, we use the datasets CN-07 and SF-07 in 2007-2008 to compare behavior differences between Chinese and the U.S. users. We discuss similarities and differences interdispersed, since often an type of behavior can exhibit both similarities and differences.

2.5.1 More calls than texts in 2007

From 2007 to 2008, calling in CN-07 and in SF-07 was the dominant mode of communication in terms of sheer numbers of communication events. However, the relative ratios of calls to texts where different in the two countries. In more detail, there are roughly 3 times more calls than texts in the daytime in CN-07, as we can see in Figure 2.2(a). To

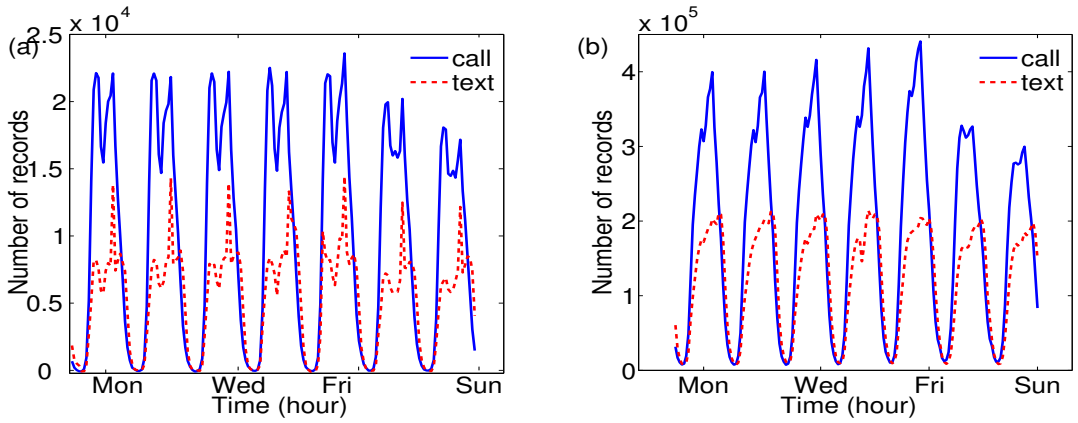


Figure 2.2: Texting and calling temporal patterns: number of calls and texts per hour over a week of (a) December 2007 in a Chinese city (b) December 2007 in Greater San Francisco Area. In (a), the numbers of calls peak at 10 a.m. and at 5 p.m. and that of texts is at 4 p.m..

quantify, an average day has 239740 calls and 52947 texts exchanged. The figure represents a typical week in December 2007 before the holidays (from the 3rd to the 10th Dec). In SF-07, an average day has 4 million calls and 3 million texts exchanged (See Figure 2.2(b)).

2.5.2 Periodicity and intra-day peaks

There are several similarities here. As expected, for both calling and texting, we discover daily and weekly periodicities: the last two days are the weekend, and they exhibit lower volumes in a typical week as shown in Figure 2.2 (a) and (b).

At the same time, there are significant differences between calling and texting regarding the intra-day behavior. In CN-07, the calling activity peaks at 10 a.m. and at 5 p.m., while the texting activity peaks at 4 p.m (See Figure 2.2(a)). The only similarity is that noon (12 p.m. - 2 p.m.) sees the lowest activities of the working hours for both texts and calls.

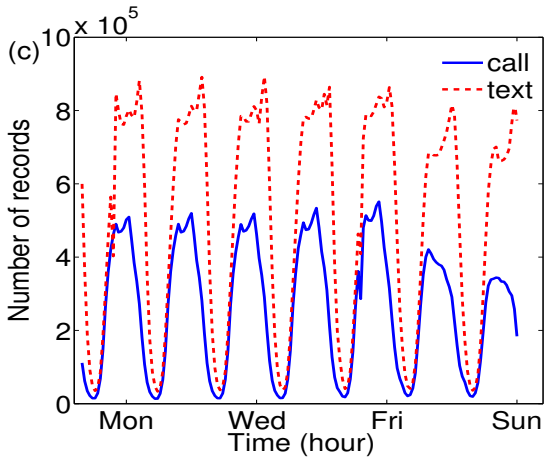


Figure 2.3: Texting and calling temporal patterns: number of calls and texts per hour over a week of (c) July 2011 in Greater San Francisco Area. In 2007, the numbers of calls peak at 10 a.m. and at 5 p.m. and that of texts is at 4 p.m.. In 2011, calls burst at 6 p.m..

In SF-07, the calling activity peaks at 6 p.m. while the texting activity has no pronounced peak (See Figure 2.2(b))! There is no obvious drop-off at lunch time. Intrigued, we investigate further and we conclude that in China, people often sleep for one hour after lunch, but this does not happen in the U.S.

Similar ratio of unique callers to unique texters per day. The number of unique callers per day is around 2.7 times that of unique texters both in China and in the U.S.. In more detail, on average there are 141K callers and 56K texters per day in China and 429K callers and 155K texters in U.S. We find this to be an interesting observation, which requires further investigation.

2.5.3 Holidays is the season for texting

As we saw, the number of callers per day is higher than that of texters in both CN-07 and in SF-07. However, as we see in Figure 2.4 (a), holidays see significantly different

user behavior: more people text than call. In CN-07, the number of texts increases by 100% and in SF-07, it also increases but only by 30%.

Moreover, we observe that the number of texts peaked on New Year's Day both in China and in the U.S. In particular, on the Spring Festival, a major holiday in China, the number of texts skyrockets, leading to a visible spike in Figure 2.4 (a). The number of texts is around five times that of calls, while the number of calls drops by 30%. This seems to suggest that, in China, texting has replaced calling as the way to spread holiday greetings.

Discussion. It seems that spikes in texting are a better indication of special events (like holidays). It would be interesting to see if this is true in emergencies or joyful events (e.g. local team winning the world cup). Understanding this, monitoring the volume of texts and calls could enable a telecom carrier detect such an event and maybe infer its nature.

2.5.4 China adopts texting earlier than the U.S.

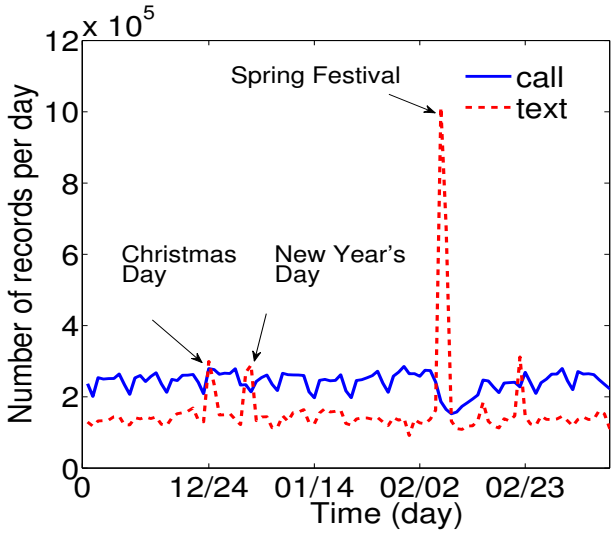
The fraction of dual-mode users is about 51% in CN-07 and 48% in SF-07. In contrast, the fraction of texters is 68% in CN-07, while this fraction is only 52% in SF-07, which is especially interesting, since San Francisco is particularly high-tech.

A possible explanation could be attributed to the policies and user contracts that the telecom industries in each country follow. In a nutshell, texting is cheaper in China, while calling is more affordable in the U.S.

Discussion: Tracing the origin of differences to culture versus money. The structure and pricing of contracts by a telecom carrier have hugh impact on communi-

cation behaviors of its customers. The Chinese carrier's policy states that a call is charged by durations with rate of 3 cents/minute anytime, even within the network of the same operator while a text is 1.5 cents/text. As a result, calling is more expensive than texting. By contrast, the U.S. operator follows the opposite policy: there is a fixed monthly fee for calling, whereas texting is charged by text message at the time frame of our SF-07 dataset. Furthermore, for several user plans, calling is free within the same operator and during off-peak hours (typically 8 p.m. - 8 a.m.), while there is no free-period for texting.

As both culture and financial considerations affect it is hard to know which factor influences more or is responsible for a particular behavior. We already saw that after-lunch naps could be responsible for the sharp daily valley among chinese users. One could also conjecture that the high spikes among chinese users during holidays could be the result of



(a) CN-07

Figure 2.4: Number of records per day with the holidays indicated. On Christmas Day and New Year's Day, the number of texts increases by 100% in the Chinese city

fixed price of texting. In the U.S., the pricing has gone through several different approaches, including the my-favourite circle, that enabled unlimited calling to a limited number of people, truly unlimited calling, pay-as you go with different pricing schemes, etc. It would be interesting to do a more extensive study to quantify the effect of pricing on user behavior.

2.6 Time evolution study

In this section, we study the evolution of texting and calling focusing on the relative growth of the two communications modes. We use the SF-07, SF-11, and SF-days datasets here. Unfortunately, we did not have a more recent dataset from the Chinese carrier to do a similar evolution study.

The number of texts surpassed that of calls with a 2:1 ratio. We find that texting has taken over in sheer number of events by flipping the number of calls over the number of texts ratio from 2:1 in SF-07 to 1:2 in SF-11. Due to space limitations, we option to show the daily snapshots that span multiple different years Figure 2.5 (a). The year 2009 seems to be the time where the number of texts overtook calls, while the number of texts in 2011 has doubled compared to that of 2007.

Note that the decrease in the number of call activity in the first two plots can be contributed to the variable number of users who are monitored. In any case, our interest is more on the relative trend of evolution between the texts and calls than on the absolute numbers.

Intrigued, we wanted to identify the causes of the increase in texting, and we find an increase in both the number of people texting as a percentage of the number of cellphone

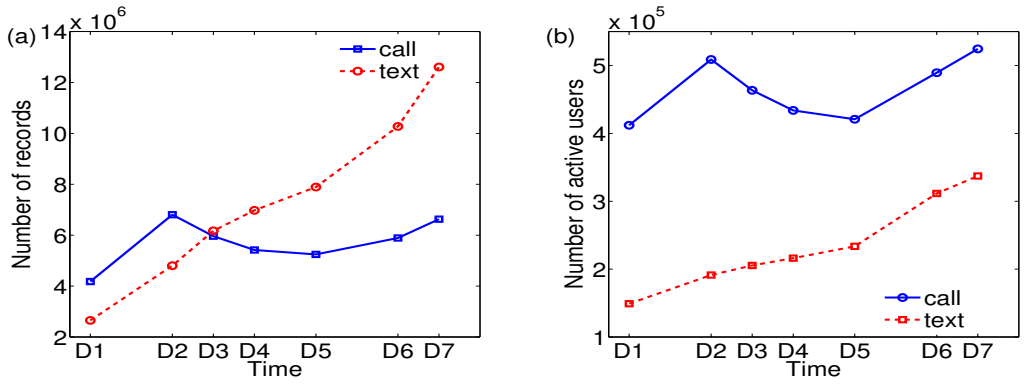


Figure 2.5: The evolution of (a) communication events, and (b) users for users active in that mode of communication over 4 years in the greater San Francisco area. The number of active texters per day surpasses that of callers in 2009. D1 = 12/08/07, D2 = 08/26/08, D3 = 03/11/09, D4 = 07/07/09, D5 = 04/14/10, D6 = 02/28/11, D7 = 07/16/11

.

users, and the number of texts per texting user.

The fraction of texters increased to 74% in 2011. More people are active in texting; as a result the fraction of texters in SF climb to 74% in 2011 from 52% in 2007. Despite the increase, the number of texters is still less than that of callers: the ratio changed from 1:3 in SF-07 to 2:3 in SF-11.

In more detail, the number of texters per day in July 2011 has doubled compared to that in December 2007, while the number of callers is only 20% (See Figure 2.5(b)). In December 2007, there are around 415,000 callers and 156,000 texters per day, while in July 2011 the two number increase to 495,000 and 319,000. At the level of a user, 65% users employ both calls and texts, 26% use only calls and 9% use only texts.

A texter sends roughly double the daily texts compared to 4 years ago. We focus on users that text actively and we plot the average number of texts per day, and show the result in Figure 2.6. On December 8th, 2007, the average number of texts

increased from 17.8 in 2007, to 37.1 in 2011. In contrast, the average number of daily calls by active callers has increased relatively little, and appears to have stayed rather constant over the last few years.

2.7 Conclusions

To the best of our knowledge, this is the first study that focuses on understanding the interplay and correlation between texting and calling of cellphone users. We use data from two cellphone carriers from China and the U.S and datasets from 2007 and 2001. We provide some definitions towards a framework for capturing multi-modal communication, to capture granularity in time and usage, with the aid of time-slices, and channels, and recommend the study at both the user and communicating pair levels. For a communicating pair, we find that the existence of texting during the weekend time-slice is the strongest indicator of persistent communication in terms of the likelihood that the pair interacts in other time-slices. In terms of cultural comparison, there are both similarities and differences. For example, holidays exhibit spikes in texts for both countries, but by a different level of increase: 100% in China, and 30% in the U.S.. Finally, our analysis suggests that the dominant status of calling no longer exists, since texting has surpassed calling in terms of absolute number of communication events by flipping the number of calls over the number of texts ratio from 2:1 in 2007 to 1:2 in 2011.

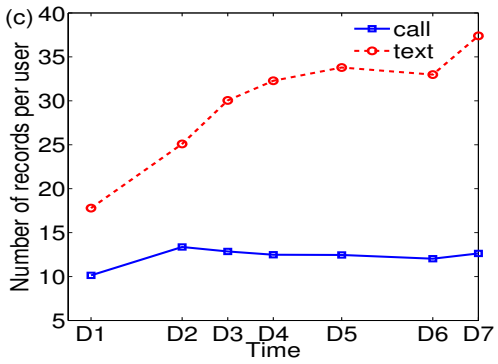


Figure 2.6: The evolution of average daily activity for users active in that mode of communication over 4 years in the greater San Francisco area. The number of active texters per day surpasses that of callers in 2009. D1 = 12/08/07, D2 = 08/26/08, D3 = 03/11/09, D4 = 07/07/09, D5 = 04/14/10, D6 = 02/28/11, D7 = 07/16/11

.

Chapter 3

Extraction of information regarding the roles of users

3.1 Introduction

Today we are seeing a diversity of available electronic modes of communication, such as e-emails and short messages, which increases the ways people can interact while at work. To facilitate the study of interactions in an enterprise, we represent a group of communicating individuals as a graph, where nodes denote employees and edges capture particchapterular communications, such as phone-calls, between two nodes. We refer to these graphs as Communication Interaction Networks, or CINs. With the ever increasing availability of CINs, it is important for enterprises to effectively analyze and understand their data and exploit any information they may infer. Such information is essential for accessing potential security risks from information leakage, as well as for improving the

communication services in the enterprise (e.g., differentiating among users based on their communication behavior or role). At the same time, understanding and baselining the different CINs today, it is a fundamental first step in identifying anomalies in the future; such as the spread of spamming or smart-phone malware.

In this chapter, we focus on three key questions:

- Q1: How do the CINs from two modern enterprises look and which are their common characteristics?
- Q2: How are the different communication modes used within an enterprise? Are there differences between ordinary employees and managers in how they use different communication modes?
- Q3: How much information can we extract regarding the roles of their participants and how can we identify hosts in different levels of the hierarchy (e.g., managers versus ordinary employees)?

We address the above questions using empirical CINs from the Enron Corporation and a communication provider. To the best of our knowledge, we are the first to present similarities between CINs from different enterprises as well as to compare CINs created by different communication modes (short messages and phone calls) within the same enterprise. In the remaining of the chapter, all our observations reflect behaviors seen in the two enterprises for which we have data. In the future, we plan to include data from more enterprises in order to identify key properties that capture intrinsic behaviors in modern corporations.

With regards to Q1, we observe that CINs from our two different enterprises have common graph structural features: they have high edge density, high clustering coefficient, and close to zero assortativity coefficient. Those similarities are present even though the CINs we study are formed using various communication modes, such as phone calls, text messages and emails. In order to further highlight these features, we compare CINs with (1) real-world complex networks and (2) synthetic random graphs with the same node and edge numbers as our CINs. That allows us to identify key structural properties of our CINs that are not observed in other graphs, even when the other graphs share the exact degree distribution.

With regards to Q2, we observe that edges in CINs formed by different communication modes, even within the same enterprise can be different. In fact, we observe that only 41% of communicating node pairs use both short messages and phone calls when interacting with each other. Moreover, we notice that employees with different roles in the enterprise have distinct communication behaviors: for example managers are more likely to utilize multiple modes to contact others whereas ordinary employees use just one.

With respect to Q3, we focus on a very specific question: *Can we distinguish between ordinary employees and managers by using only the graph topology?* Towards this end, we propose **HumanRank** (§3.6.1), a method of ranking individuals based on their position in the hierarchy (e.g., CEOs having higher rank than ordinary employees). Next, using our HumanRank method, we introduce an unsupervised and parameter-free algorithm that separates managers from ordinary employees. Our algorithm achieves above 70% accuracy in labeling managers and employees. Especially, our method is better in capturing

managers and outperforms the state-of-the-art [1] by more than 15%.

The chapter is organized as follows. In §5.2, we review related work and discuss how we differ from it. In §5.3, we describe the data sets and the metrics we used for analyzing CINs. In §3.4 and §3.5, we present the analysis of CINs and compare with other real-word and synthetic networks. In §3.6, we present HumanRank and compare it with others. We conclude the chapter in §5.6.

3.2 Related Work

Analyzing complex networks. Real-world complex networks have attracted significant attention over the past years. Such networks include: scientific collaboration networks [10], actor collaboration networks [11], online social networks [**?**][12], the Internet AS-level topology [13], the Web graph [11], and biological networks [14]. In this work, we capitalize on the methodologies and the different graph metrics proposed in the literature and use them to shed light to a different kind of complex network, namely, the interaction communication networks within an enterprise. As we discuss next in more detail, we analyze our graphs using a diversity of metrics proposed in previous work, such as: degree and degree correlations [15] [16], clustering coefficient [17], assortativity [18], path length, betweenness [19], and network motifs [20].

Ranking nodes and identifying roles in networks. Significant information can be extracted from structures of real-world complex networks. Kleinberg shows how to extracts hub and authority pages from the structures of WWW [21]. Google uses the PageRank algorithm to assign importance values to web-pages [22]. Our method, is inspired

by PageRank, but aims to answer a very different question. That is, with HumanRank we identify the roles of employee in an enterprise, e.g., managers, CEO, etc. by looking at the structures of their communication networks. In similar spirit, Karagiannis et al. also study email exchanges in a large enterprise [23], but their work focuses on profiling users based on personal communication features, such as the volume of emails being send and the number of email addresses in the contact list of a user, and not the structure of the network.

Inferring hierarchy in enterprise CINs. Here we describe the studies that are more related to ours. Using the same Enron dataset with us, Shetty et al. propose an entropy-based method to rank employees based on personal communication patterns [1]. As we show in §3.6, in our experiments HumanRank performed better than the entropy-based approach [1]. Powe et al. [24] propose a method that uses topological features to rank employees, such as the degree of the nodes. The method in [24] requires significant input from the user, such as the assignment of weights to features, which is hard to do in practice. By applying this method to our data and using the same weight for all metrics we observed poor ranking results. Because of the low quality of the results we do not include them here. Gupte et al. propose a method to identify social ranks in online social networks [25]. Social networks are based on explicit friendship relationships, which is different from the exchange of message as we use them for CINs. For example, in a social network we expect people to interact with people of the same status, whereas in CINs we expect managers to communicate more frequently with their employees. Maiya et al. proposed a method to infer the tree-like hierarchy in the two networks of the George W. Bush and the Barack H. Obama administrations [26]. There work is based on strong assumptions about the

underlining shape of the graph and relies heavily on information other than the structures of the graph.

3.3 Datasets and Graph Metrics

3.3.1 Datasets

For facilitating the analysis of communication patterns, we represent a group of people as a network $G(V, E)$ (or graph), where nodes (V) represent individuals and edges (E) capture particular interactions. Here, we study enterprise CINs, where nodes are employees in the same enterprise and edges are formed from the chapter exchange of phone-call, short message or emails among them. We will refer to these CINs as CALL, SMS and EMAIL networks respectively.

CALL and SMS: These data capture communications among 235 employees from a city branch of a large corporation in 2007. For each employee, we have their job titles: manager and ordinary employee. We use this information as ground truth when evaluating our algorithms. The data span over half a year. The SMS data is from the exchanges of short messages from the employees' cell-phones. The CALL captures all the phone calls between the same set of employees. For CALL, we have callers and callees and thus we produce directed edges. Unfortunately, SMS is undirected because sender and receiver are not able to be distinguished in the dataset. Edge weight is defined as the number of communications between two participants during the data span.

EMAIL: This network is extracted from the Enron public email dataset collected by J. Shetty [1]. The public data is comprised of email exchanges among 151 employees that span over 2.5 years. For 101 employees, we have his/her name job title (e.g., CEO, manager) and job description. The job title is unknown for 50 Enron employees. We always include these nodes in our graphs, but we do not use them when we evaluate our methods. Like CALL, EMAIL is directed. However, in order to study CALL, SMS and EMAIL in a uniform way, we consider them as undirected networks except in motif discovery.

To the best of our knowledge, this work is the first that focuses on enterprise CINs, using data from two different enterprises and three different communication channels. The enterprises we study are from two countries with different cultures, which adds to the diversity of our data. For example, in 2007 short message in China is a more popular communication means than in U.S. from our another study. Moreover, studying CINs from different communication channels allows more detailed analysis of the behaviors within an enterprise. For example, the interactions using phone-calls appears to be more direct than contacts by short message and email. This can explain the higher similarity of SMS and EMAIL CINs compared to the CALL graph, as we show in §3.4 in more detail. The collection of data from more enterprise organizations can further support our findings, and this is something we plan to pursue in the future. However, we need to underscore that this is not an easy task due to security and privacy concerns.

3.3.2 Metrics for analyzing and comparing graphs

Degree of nodes. We begin our analysis of enterprise CINs by looking at the

degrees of their nodes. Given a network $G(V, E)$, V denotes the set of nodes and E the set of edges. The $|.|$ notation denotes the cardinality of a set. The undirected degree of a node u, denoted with k_u, is the number of its direct neighbors. The average degree of a graph is given by $\bar{k} = \sum_{u \in V} k_u / |V|$. The edge density (ED) is given by $2 \cdot |E| / (|V| \cdot (|V| - 1))$. The degree distribution is the probability distribution of those degrees over the entire network, and is denoted as $P(k)$. It is often captured with a complementary cumulative distribution function (CCDF), $CCDF(k) = \sum_{k}^{|V|} P(k)$, and plotted in a logarithmic scale to highlight if there is a power-law distribution [16].

Degree-to-degree correlations. We also examine the degree-to-degree correlations in the graph. This value is a correlation between all nodes with degree k and the mean degree of their direct neighbors. These degree correlations are often summarized in a single value called assortativity (r), which is expressed as:

$$r = \frac{|E|^{-1} \sum_i k_u^i k_v^i - [|E|^{-1} \sum_i \frac{1}{2}(k_u^i + k_v^i)]^2}{|E|^{-1} \sum_i \frac{1}{2}((k_u^i)^2 + (k_v^i)^2) - [|E|^{-1} \sum_i \frac{1}{2}(k_u^i + k_v^i)]^2}$$

where k_u^i, k_v^i are degrees of two end-points (u, v) of the ith edge, with $i = 1...|E|$. When $r < 0$, it indicates that high degree nodes in the network tend to link with low degree nodes and when $r > 0$ high degree nodes like to connect to high degree ones [18]. The assortativity of a random network is near zero, which indicates there is no degree-to-degree correlation.

Distances between nodes. We measure how far two randomly selected nodes in a graph are by using the distance distribution (a.k.a path length distribution). For that metric, we first calculate the length of all pair shortest paths in the graphs and plot their distribution. The diameter of a network is defined to be the longest path, and is often used as a metric for comparing graphs [27, 28]. The number of all pair shortest paths that pass

through a node measures the centrality of the node [19].

Graph decomposition. A powerful set of metrics for analyzing graphs are based on decomposing the graph and studying it as a set of small sub-graphs. Here, we are interested in maximal cliques [29]. A maximal clique is a clique that cannot be extended by including one more adjacent node, and typically reflects a group of nodes having a close relationships with each other [29]. The clustering coefficient of a node is defined to be the ratio of the number of existing links over the number of possible links between its neighbors [17]. Given a network $G(V, E)$, the clustering coefficient of a node $C(u)$, $u \in V$ is:

$$C(u_i) = \frac{2|\{e_{xy}\}|}{k_i(k_i - 1)}$$

where u_x,u_y are neighbors of u_i. The value of $C(u)$ captures the probability that two u's neighbors are also connected. This is often used as a measure of how well connected the local neighborhoods are in a graph. The average clustering coefficient of a graph is given by $\bar{C} = \sum_{u \in V} C(u)/|V|$.

3.3.3 Generating synthetic random networks for comparison

In order to highlight the unique structural properties of real-world CINs, we generate various synthetic graphs for comparison. Intuitively, using synthetic data we can generate networks that have exactly the same sizes (number of nodes), or even networks that follow identical degree distributions with our CINs. Here we use two ways to produce a synthetic network: (a) one is to depend on an existing real network, randomly re-arrange

edges and get a randomized version of the network and (b) the other is to rely on a graph generation model and produce a network with a predefined set of characteristics.

NEP: To generate a random network with the same number of nodes and edges as our initial enterprise CINs, we use the 0k-preserving re-wiring technique introduced in [15]. In a nutshell, at each iteration of the re-wiring process, one edge is selected at random and then moved to a random location, that is, it connects a different pair of nodes. This step is repeated multiple times to give a random version of the initial graph. Here, we executed $10 \ |E|$ re-wirings, as suggested in [15]. In the chapter, we call random networks produced by this process as Node and Edge Preserving (NEP) graphs.

DDP: For generating synthetic graphs with exactly the same degree distribution as our CINs, we used the 1k- preserving re-wiring technique from [15]. In this technique, two edges are chosen at random and then exchange node connections, that is, two given edges (v, w) and (m, n) are turned into two new ones (v, n) and (m, w). Here, we executed $10 \ |E|$ re-wirings, as suggested by the work [15]. The main advantage here is that the degree distribution of the input network is exactly preserved. We use the name Degree Distribution Preserver Network (DDP) to refer to this type of synthetic networks.

BA: Here we generate synthetic graphs that follow a power-law distribution and have the same number of nodes as our CINs. Towards this end, we use the well-known Barabasi−Albert (BA) [11] generator. The generator uses the concept of preferential attachment. It keeps adding nodes with each new node preferring to connect to nodes with high degree[11]. Since we want the graph to follow a power-law distribution, we do not impose any restriction on the resulting number of edges. That is, the final graph will have

the same nodes as the target CIN but potentially different number of edges.

R-MAT: Finally, we use the R-MAT generator to make synthetic graphs (a) with the exactly same numbers of nodes and edges as our initial CINs and (b) that follow the structure of other popular complex networks, such as the Web graph[30]. In a nutshell, the generator uses an iterative process based on fractals [30] resulting in self-similar graphs. The model generates directed graphs, which enables a detailed comparison with our directed CINs.

3.4 Characterizing Enterprise CINs

Our goal in this section is to answer the following questions: (1) *How do the three enterprise CINs look and what are their differences and similarities*, (2) *How different are the CINs from popular complex networks?* and (3) *How different are the CINs from synthetic graphs we generated?*

3.4.1 Comparing the CALL, SMS, and EMAIL CINs

We summarize a set of scalar metrics for the three CINs in Table 3.1, and plot six non-scalar metrics in Figure 3.1. As we see from Table 3.1, all three networks are of similar scale, having 235, 234, and 151 nodes for the CALL SMS, and EMAIL CINs respectively. Even though the number of nodes is small, relatively to other complex networks, their number of edges is high leading to high edge densities ($> 12\%$). The degree distribution for all three graphs is shown in Figure 3.1(a). Even though there are some differences, all three CINs follow similar trends in their degree distributions.

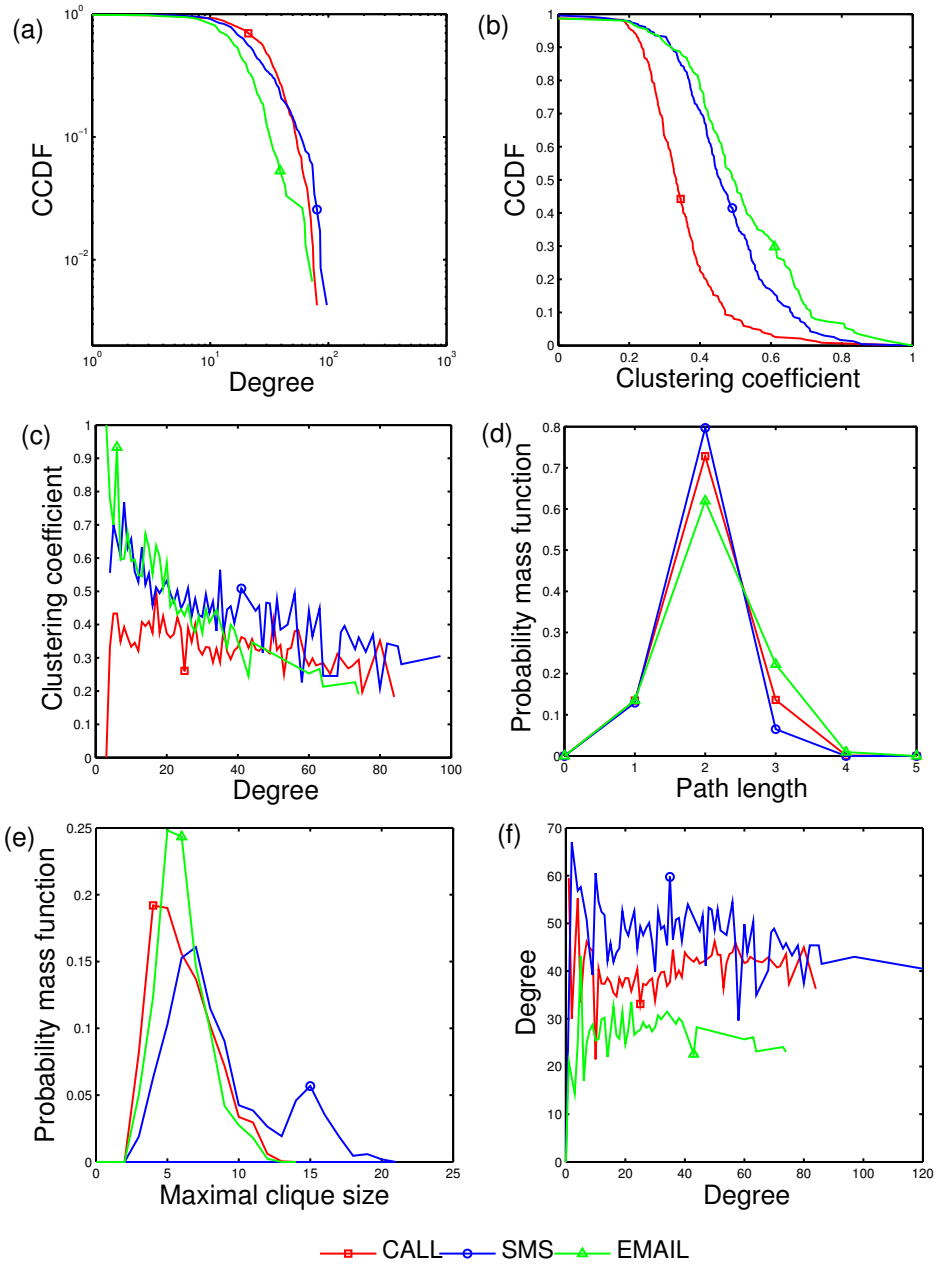


Figure 3.1: Comparison of the CALL, SMS, and EMAIL CINs using six different no-scalar metrics. All networks have similarities in their degree distributions, clustering coefficient, path lengths, and their degree-to-degree correlations. The SMS and EMAIL are more similar to each than to the CALL graph.

Our next step is to compare the clustering coefficient in the graphs. In Table 3.1, we see that CALL's average clustering coefficient is the lowest. Even though CALL and SMS are produced by the same group of employees, they have difference, indicating individuals have different communication behaviors depending on the communication channel they use. To highlight the differences, we show the distribution of clustering coefficient over all nodes in Figure 3.1(b). We see that more than 50% of the nodes in SMS and EMAIL have clustering coefficient larger than 0.5, whereas the percentage of such nodes in CALL is only 7%. This figure shows that SMS and EMAIL CINs are more similar to each other than to the CALL graph. In Figure 3.1(c), we show the relation between the clustering coefficient and the degree of nodes in the graphs. The x-axis shows the degree and the y-axis the mean clustering coefficient for all the nodes having that particular degree. We see that all graphs show a similar pattern, where the clustering coefficient decreases as the degree of the node increases. Moreover, we see again the similarity between the SMS and the EMAIL CINs. We plot the distribution of maximal cliques of various sizes in Figure 3.1 (e). In EMAIL and SMS networks, there are more maximum cliques with more nodes than in the CALL network. We speculate that we observe this because short messages and emails may be sent between a set of people with the option to reply to the entire group (e.g., reply-all in emails), but this cannot be done with phone-calls.

The degree-to-degree correlation is captured in the assortativity coefficient shown in Table 3.1. Interestingly, the three values are around zero with values 0.095, -0.084 and -0.061 for the CALL, SMS and EMAIL respectively. To study degree correlations in more detail, we plot the average neighbor degree metric in Figure 3.1(f). The x-axis shows the

Table 3.1: Scalar graph metrics for CALL, SMS and EMAIL CINs compared to real and synthetic Networks. $|V|$: number of nodes, ED: Edge density, $\bar{k}$: average degree, D: Diameter. P: Path length and $\bar{C}$: average clustering coefficient

ID	Family	Data set	$\|V\|$	ED	$\bar{k}$	r	D	P	$\bar{C}$
1	CINs	CALL	235	13.4%	31.5	0.095	4	2	0.35
2		SMS	234	12.9%	30.1	-0.084	3	1.9	0.47
3		EMAIL	151	13.4%	20.2	-0.061	4	2.6	0.51
4	Networks	Physics [10]	52,909	0.018%	9.7	0.363	20	5.9	0.43
5	Capturing	Biology [10]	1,520,251	0.0009%	14.8	0.127	21	4.4	0.072
6	Human	Actor [17]	282,219	0.027%	78.6	0.22	16	3.65	0.63
7	Relationships	Cyworld [28]	12,048,186	0.0002%	31.6	-0.13	18	3.2	0.16
8		MySpace [28]	100,000	0.14%	137.1	0.43	8	2.7	0.43
9	Other	Internet [31]	35,186	0.012%	4.2	-0.22	9	3.83	0.29
10	Networks	WWW[11]	269,504	0.003%	9.19	-0.05	40	7.53	0.24
11		Power grid[17]	4,941	0.054%	2.67	0.003	46	18.7	0.080
12		Protein[14]	1,870	0.13%	2.35	-0.161	19	3.75	0.067
13		Neural[17]	297	4.9%	14.5	-0.163	5	2.45	0.29
14	Synthetic	NEP	235	13.4%	31.5	-0.008	3	1.88	0.13
15	Networks	DDP	235	13.4%	31.5	0.009	4	1.95	0.2
16	(see §3.3.3)	R-MAT	234	13.2%	31	0.123	3	1.93	0.21
17		BA	235	1.4%	3.16	-0.15	8	3.8	0.03

degree of a node and the y-axis shows the average degree of all the neighbors of all nodes with the particular degree. Different from other complex networks, like popular OSNs, the degree of a node in CINs does not reveal information about the degrees of its direct neighbors. In other words, the average degree of the neighbors of a node u, is close to the

average degree of the entire graph, and is not correlated with the degree of u. At first, this might suggest that CINs are like random networks, but it does not hold because random networks tend to have low clustering coefficient (lower than 0.15).

Next, we measure the network diameter and average path length over all pairs of nodes. From Table 3.1, we observe that the diameter is 4 and average path length is close to 2, indicating that our graphs are small-world and the majority of nodes are just two hops away. The path length distribution is shown in Figure 3.1(d). We see that all three graphs share very similar path length characteristics.

3.4.2 Comparing CINs with popular complex networks

The comparison with other real-world networks is essential for understanding the unique features of CINs. From our study, we observe that CINs are different from other networks in the following: they have high edge density and clustering coefficient, and close to zero assortativity. To facilitate our comparison study, we collect five networks that capture some form of social interactions and five networks that do not represent social relationships. We then calculate our metrics of interest and compare them with the three CINs. We summarize our set of scalar metrics for all graphs in Table 3.1. All the graphs we use for comparison are publicly available and a detailed description is provided in their corresponding citations (see Table 3.1). Next, we present our findings in more detail.

From the networks we used for comparison, those representing human relationships (IDs 4-8 in Table 3.1) have lower edge density than our studied CINs. In addition, the three CINs have close to zero assortativity ($|r| < 0.1$) and high clustering coefficient ($\bar{C} > 0.35$).

This is different from in networks with IDs 4-8, where either the $\bar{C}$ is high or the $|r|$ is low, but never both at the same time. For example, the Actor network (ID6) has high $\bar{C}$, but its assortativity is positive; showing that high degree nodes tend to connect with other high degree nodes and low degree nodes with others with low degree. Similarly, the Cyworld network (ID7), has low assortativity ($|r| = 0.13$) but its clustering coefficient is also small ($\bar{C} = 0.16$). Our results indicate when human relationship networks have high clustering, they also have high absolute values in their assorativities.

Our three CINs are also different from the group of networks with IDs 9-13. All these networks, with the notable exception of ID13 (Neural network), have lower edge density than our CINs. The ID13 network is of the same scale as the three CINs and it also has relatively high edge density. However, its $|r| = 0.163$ value is higher and its average degree ($\bar{k} = 14.5$) is lower than three CINs.

The number of nodes and edges of a graph can affect other graph metrics. We see this in Table 3.1, where the diameter and average path length of the three CINs, it is much lower than all graphs with IDs 4-12. Intuitively, in a dense network with few nodes, is much more likely to have a short diameter and short paths than in a large and sparse network, such as the Web graph (ID10). It is therefore important to compare our graphs with other networks of the same size (number of nodes and edges), which is what we discuss next.

3.4.3 Comparison with synthetic networks

Here we have two goals. First, we want to compare our CINs with graphs of the same scale (e.g., same number of nodes). Second, we want to see which features of CINs

are not likely to happen by chance. Such features can then be traced back to the driving forces that lead to their formation. Details on our synthetic networks are given in §5.3. For all our experiments, we have generated 10 random versions for each synthetic graph. Here we include results from a single run, since all versions resulted in similar results.

We have generated synthetic random networks with characteristics from all three CINs. We observed qualitatively similar observations using all three networks. For ease of exposition of the findings we only include our results from the CALL network.

We use four synthetic networks for comparison. The NEP and DPP are based on re-wiring of the CALL network, and the BA and R-MAT are produced by graph generators (see §5.3 for details). For each synthetic network, their scalar metrics are listed in Table 3.1 and we plot their non-scalar metrics in Figure 3.2.

From Table 3.1, we see that our graphs are very different from the BA power-law network. The BA network has few edges (only 372) and small clustering coefficient (0.03). The difference with BA is highlighted in the degree distribution of Figure 3.2 (a). Moreover, the difference with BA is clearly observed in all plots of Figure 3.2, where the values of BA are always very far away from those of the rest of the graphs.

The R-MAT graph is closer to the targeted CINs compared to the other synthetic networks. We see the similarity in all the scalar metrics in Table 3.1. From Figure 3.2, we see that R-MAT follows the CALL graph very closely in all metrics except the clustering coefficient. This shows that the high clustering coefficient observed in CINs is unique and hard to be captured by the R-MAT generator. We see this also in Figure 3.2 (e) where R-MAT is not able to generate the large maximal cliques observed in the CALL network.

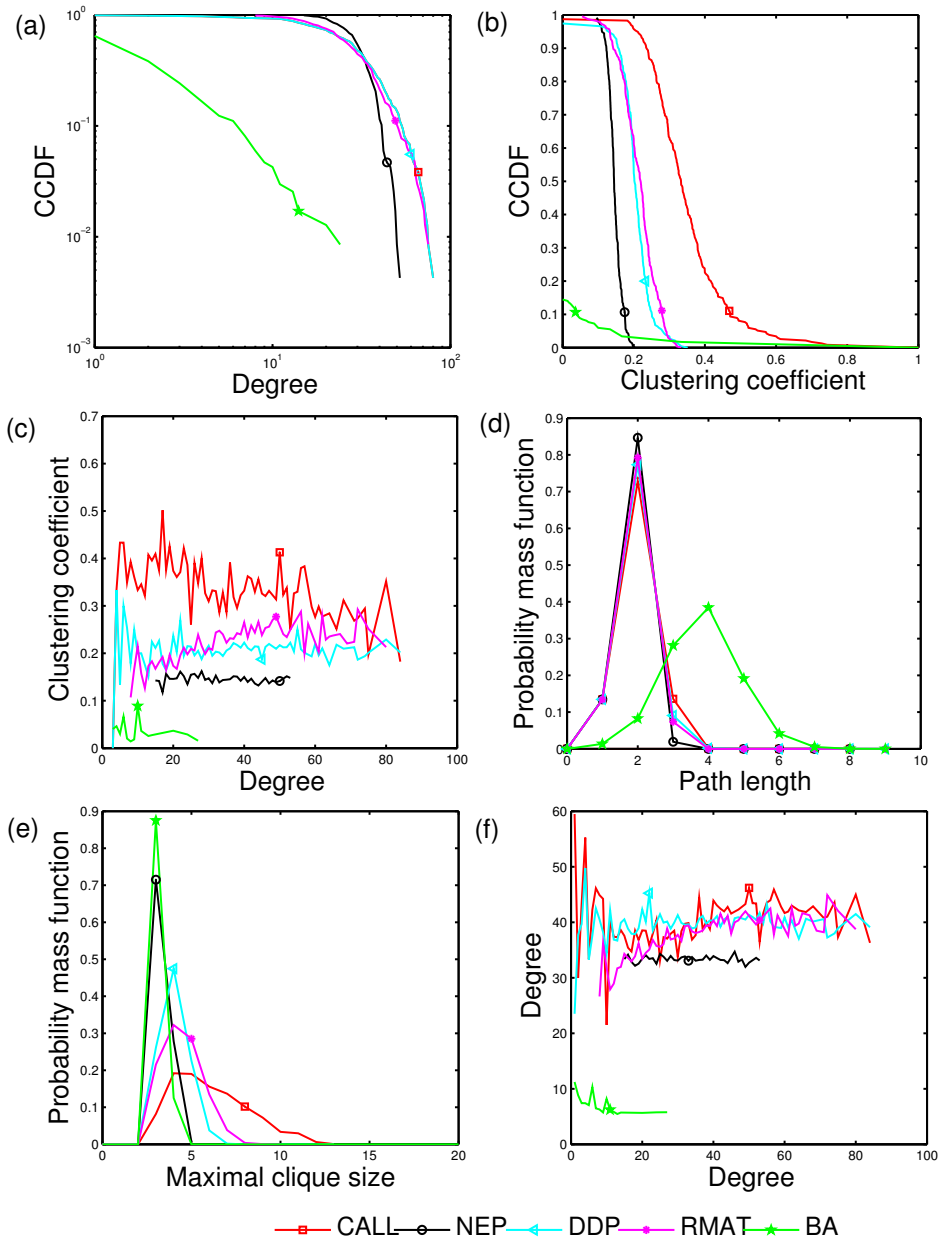
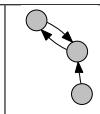
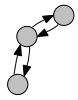
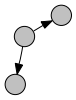


Figure 3.2: Comparing the CALL network with four synthetic graphs. The CALL CIN is different from random graphs, especially in the clustering and the formation of large cliques. From all the synthetic graphs, the power-law (BA) is the most different from our CALL CIN.

Table 3.2: Presentation of selected directed motifs of size three.

				
Motif ID	1	2	3	4
CALL	21.4%	17.2%	6.9%	3%
NEP	1.6%	0.1%	26.6%	0.004%
DPP	1.1%	0.06%	30.2%	0.001%
R-MAT	9.8%	1.5%	17.6%	0.09%

Our enterprise CINs are not random and they show properties that are unlikely to occur by the random connection of nodes. In particular, the degree distribution and clustering coefficient of the CALL CIN cannot be attributed to the high edge density. We show this in our comparison with the NEP randomized graph version. This synthetic network has exactly the same number of nodes and edge density as our CIN, but as we see from Figure 3.2 (a) and (b) its degree distribution and clustering coefficient are different.

We compare the CALL graph with a DPP random graph that has exactly the same degree distribution. Even though we see from Figure 3.2(a) that the degree of DPP and CALL are identical, the clustering coefficient shown in Figure 3.2(b) is very different. Moreover, the randomized process of generating DPP fails to form the large cliques we see in the CALL network, as we observe from Figure 3.2(e). On the other hand, the NEP and DDP random networks have degree-to-degree correlations (e.g., assortativity) and path lengths (Figure 3.2(d)) that are very similar to the CALL graph's.

Our findings indicate that the clustering coefficient of the CIN graphs is the main property that distinguishes them from the synthetic graphs. To further highlight these differences, we decompose these networks to their corresponding sub-graphs (a.k.a. motifs [20]) of size three. We summarize motifs we found particularly interesting in Table 3.2. Motifs 1,2 and 4 occur with much higher frequency in the CALL graph, compared to the synthetic graphs. In particular, the directed full-way clique-motif (4) appears two orders of magnitude more often than in any of the synthetic networks. This reflects the nature of CALL networks that edges represent human interactions, so interactions are usually bi-directional. For example, the motif 3 shows two edges that are not bidirectional. This motif happens very frequently by chance, but in CINs its frequency is considerably lower. The findings hold in all three CINs.

3.5 Communication patterns in CALL and SMS

A unique advantage of our dataset from the communication provider is that we have two different CINs, one using short messages and one using phone calls, for the same group of employees. This enables us to study the behavior of individual employees when they use those two communication modes. In this section we answer the following questions: (a) *Are the local graph features (e.g., the degree) of an employee in CALL similar to his or her corresponding characteristics in SMS?*, (b) *Are the employees likely to use single-mode or multi-mode communications?* and (c) *Are there any differences between ordinary employees and managers?*

To answer the first question, we use the scatter-plots in Figures 3.3 (a) and (b)

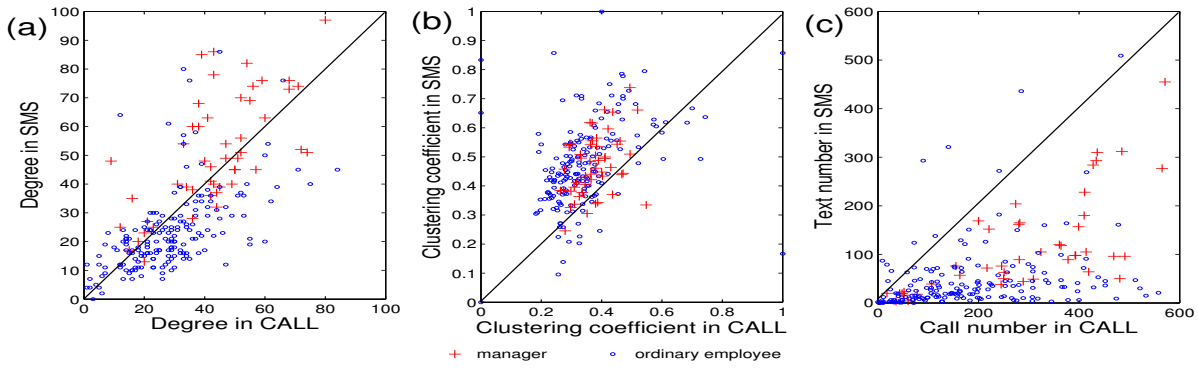


Figure 3.3: In the scatter plots we compare the characteristics of each employee in the communication provider's enterprise using the CALL (x-axis) and SMS (y-axis) CINs. In (a) and (b) we compare the degrees and clustering coefficients for the managers (crosses) and ordinary employees (dots), for the CALL and SMS networks respectively. In (c) we compare the number of phone calls made and short messages send by each employee over the duration of our dataset.

where we compare the degree and clustering coefficient for each employee in the two graphs. In both scatter-plots we see that some points are far away from the diagonal, showing differences in how the two communication modes are used by some employees. On the other hand, as a general trend, both plots indicate a loosely positive correlation, where if a node has high node degree or clustering coefficient in one network, it will have high degree of clustering coefficient in the other network. This supports our observation in the previous section where the two CINs showed similar structural graph properties.

An interesting observation from Figure 3.3 (a) is that 40% of the points are above and 60% below the diagonal. This shows that most employees have a larger local neighborhood size in the CALL compared to the SMS graph. On the other hand, the clustering coefficient scatter-plot (Figure 3.3 (b)) shows a different trend, where 85% of the points are now above the diagonal. This shows that even though most employees have smaller degree

in the SMS CIN their local neighborhood is on average better connected. This observation is suggestive of a collaborative behavior where employees tend to use short messages to communicate with their close collaborators forming well connected neighborhoods. In the next section, we show results where the higher clustering coefficient of the SMS graph can explain why the SMS CIN is better for inferring hierarchies compared to CALL.

For the second question, we find 2056 communication node pairs to be shared by both CALL and SMS, while 1634 and 1468 communication node pairs exclusively belong to CALL or SMS respectively. Our findings are graphically illustrated in Figure 3.4. From the figure we see that the corresponding percentages are 41%, 31% and 28%, which are very close to each other. In other words, if we randomly select a pair of communicating employees, is equally likely to use only short message, only phone calls or both. Interestingly, it is more popular in this enterprise for employees to use a singe-mode than a multi-mode communication. This observation indicates that when we study interactions within an enterprise adding more communication modes increases the number of observed links in the interaction graphs.

For the third question, we plot total number of calls made by an employee against his or her total number of short messages. We illustrate these results in the scatter-plot of Figure 3.3 (c). We see that most points are below the diagonal, showing that employees make more phone calls than text messages. It is very interesting to observe that managers are closer to the diagonal line than ordinary employees. This indicates that managers use both phone call and text to contact others in a relatively balanced way. In contrast, ordinary employees are inclined to use phone call as their priority choice. These observations motivate

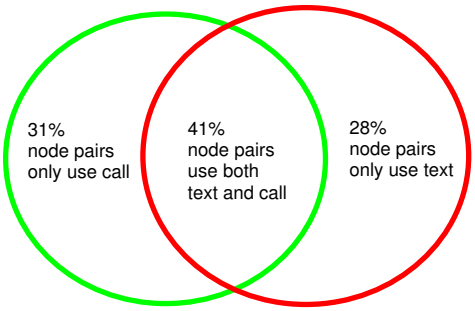


Figure 3.4: A Venn diagram comparing the number of communicating pair of nodes that use: (a) only short messages, (b) only phone calls, and (c) both short messages and phone calls. From the diagram we see than most of the communicating node pairs use a single-mode communication and only 41% are observed to use multi-mode communications.

our next section where we utilize the information within CINs in order to infer the roles of different employees.

3.6 Detecting Hierarchy Structure in Enterprise CINs

chapterIn this section, our goal is to use CINs to infer the position of employees in the hierarchy of the enterprise. Moreover, we want to see if the choice of CIN (e.g., CALL versus SMS) can affect the inference process.

3.6.1 HumanRank: Ranking employees in enterprises using CINs

Any ranking method works as follows. First, it assigns a score to each node (employee) in the graph (enterprise). Then, it ranks all nodes by sorting them according to their assigned scores. The node with the highest score gets rank=1 and the one with the lowest gets $|V|$. In our case, our goal is to give a score to each node that describes its

status in the enterprise.

Our ranking method is motivated by the following observation: The importance of an individual can be inferred from the importance of the people he or she interacts with. Intuitively, senior managers will interact more with other managers than ordinary managers. A similar observation regarding the structure of the Web lead to Google's PageRank [22]. In PageRank, if a web page is linked from other important web pages, it is regarded as a more important page.

In this section, we introduce **HumanRank**: An iterative process of ranking individuals based on their position in the hierarchy. We have adapted PageRank to fit the requirements of our problem. The PageRank algorithm treats edges differently based on their direction and one of its simplied versions for calculating v's PageRank is defined below:

$$PR(v) = \sum_{w \in B(v)} PR(w)/D(w)$$

where $B(v)$ contains all pages linking to page v and $D(w)$ is the outdegree of w.

In our work, we treat a CIN as an undirected network. In the Web, if a small website w has a hyperlink to a popular website (e.g., google.com), this does not say anything about the importance of website w. However, if an individual contacts the CEO of a company, there is a high chance that this individual is also important. We also apply PageRank without our adaptation and observed lower accuracy in all our directed graphs. In HumanRank the hierarchy score of a node is defined as follows:

$$H(v) = \sum_{w \in L(v)} H(w)$$

Where $L(v)$ is the set of v's neighbor nodes. The complete process is outlined in Algorithm 1. In practice, the iteration process stops when the score of each node does not change significantly from the previous step. In all our experiments, we used a 1% threshold, which resulted in convergence after 10 to 20 iterations

Algorithm 1 HumanRank

Require: a network G

for each $v \in V(G)$ **do**

$H(v) \leftarrow 1$

end for

while $H(v)$ changes **do**

for each $v \in V(G)$ **do**

$NH(v) \leftarrow \sum_{w \in L(v)} H(w)$, where w is one of v's neighbor nodes

end for

$\overline{NH} \leftarrow \sum_{v \in V(G)} NH(v)/|V(G)|$

for each $v \in V(G)$ **do**

$H(v) \leftarrow NH(v)/\overline{NH}$

end for

end while

return $H(v), v \in V(G)$

Exploring other ranking methods. We are interested in the following ques-

tion: *Why not use a simple graph metric for ranking, such as the degree or the clustering coefficient of a node?* As we see from Figure 3.3 (a), the degree of a manager is not absolutely higher than an ordinary employee, especially in the CALL network where the average degree of ordinary managers (43.9) is higher than that of senior managers (38.4). The clustering coefficient also shows its limitations. These observations highlight the difficulty of the problem at hand and suggest that trivial solutions are unlikely to lead to good results. We revisit the comparison with other methods in §3.6.3.

3.6.2 Hierarchy detection using HumanRank

Given the ranking results from HumanRank, we want to form two groups. Intuitively, nodes with low HumanRank scores are more likely to be ordinary employees, whereas those with high scores are more likely to be managers. In addition, we want our classifier to be unsupervised. That is, we do not want to depend on training data, which are often hard to obtain in practice. We compare our approach with supervised machine learning classifiers in §3.6.3.

The steps of our classification algorithm are outlined in Figure 3.5. First, we calculate HumanRank on the input CIN. Next, we use K-means to group nodes into two disjoint groups based on their HumanRank score. The last step is to label each cluster as ordinary employees or managers. We use the average ranking score of each group to achieve this. All the nodes in the cluster with the high score are labeled as managers. Similarly, all the nodes in the other cluster are labeled as ordinary employees.

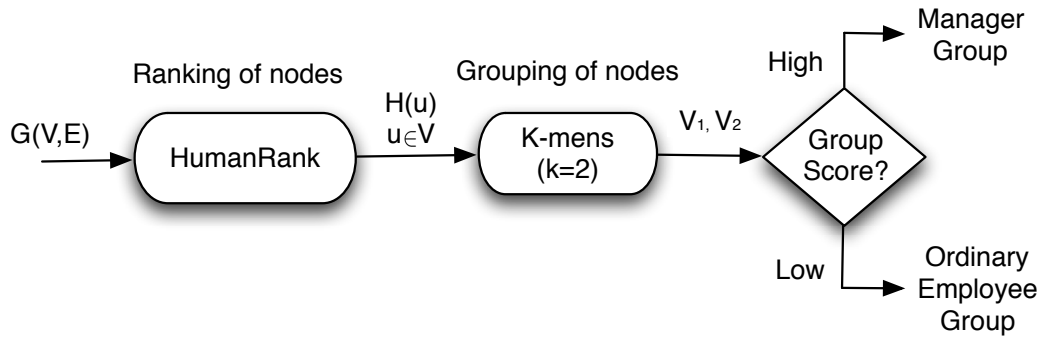


Figure 3.5: The two step process for labeling nodes as ordinary employees or managers. All the nodes are firstly ranked according to HumanRank and then are assigned in two groups using k-means (k=2).

3.6.3 Experimental Evaluation

In our dataset, we have the job description of every employee, therefore we use this information as **ground truth** to evaluate our algorithms. The classification **accuracy** is defined as the number of correctly classified employees divided by the total number of employees with ground truth. **F-score** is a measure of a test's accuracy and considers both precision and recall.

Ranking results

For evaluating the ranking results, and comparing it with other methods, we use the following test. We rank nodes using their score by a given method, e.g., HumanRank. We then measure how many managers are located in the top k ranked nodes and measure this for different k's in the range $1...|V|$. We compare our results with the state-of-the-art in hierarchy detection, which uses graph entropy [1]. In addition, we compared with ranking based on the node degree and PageRank. We also used ranking on clustering coefficient

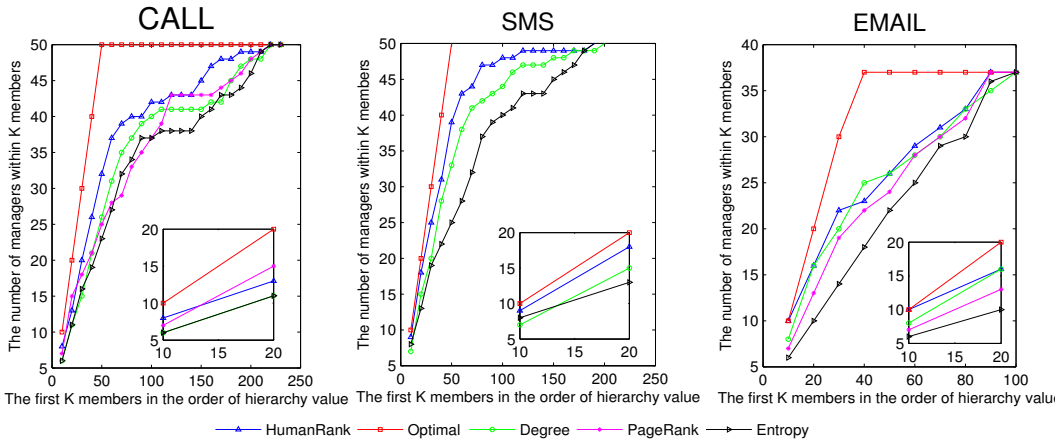


Figure 3.6: Ranking results: The plots show the number of managers identified in the top k nodes, as ranked by our HumanRank and three other ranking methods. We also compare our results with an optimal ranker where all managers have higher score than ordinary employees. PageRank is not applied on SMS because SMS is undirected.

that did not give good results and we omit it from the remaining of the chapter.

The comparison results are illustrated in Figure 3.6. We also report the results of an Optimal ranking method, which always gives better ranking to managers than ordinary employees. As we see from the figure, the HumanRank gives results that are closer to the optimal ranker than others. In particular, we are doing significantly better than the state-of-the-art method, which uses entropy [1]. In Figure 3.6, we magnify the ranges up to rank 20 for better comparison. In the top 10 ranks for all three CINs, our approach gives eight, nine, and ten managers, which corresponds to 90% of the time on average. This is in contrast to 70%, 65%, and 67% for the degree, PageRank, and entropy rankings, respectively.

It is also interesting to observe that the ranking results are better when we use the SMS network compared to the CALL network, even though they capture interactions

from the same enterprise. This suggest that the "informal" exchange of SMS can reveal the hierarchical structure of an enterprise better than the more "formal" exchange of phone-call. In the EMAIL network, the ranking appears more challenging for all methods. However, even with this data set, our algorithm gives better results over a long range of ks. We hypothesize that the lower accuracy in the Enron dataset is due to the removal of some email exchanges after personal requests by the participating individuals [32]. Since retrieving the lost emails is hard, establishing causality can be difficult.

Hierarchy detection

We group the competing methods into supervised and unsupervised solutions. Supervised solutions combine graph features for different nodes. We try different machine learning methods: logistic regression, SVM, Random forest, and Bayesian Networks. The features we used for classification are: degree, clustering coefficient, betweenness, average

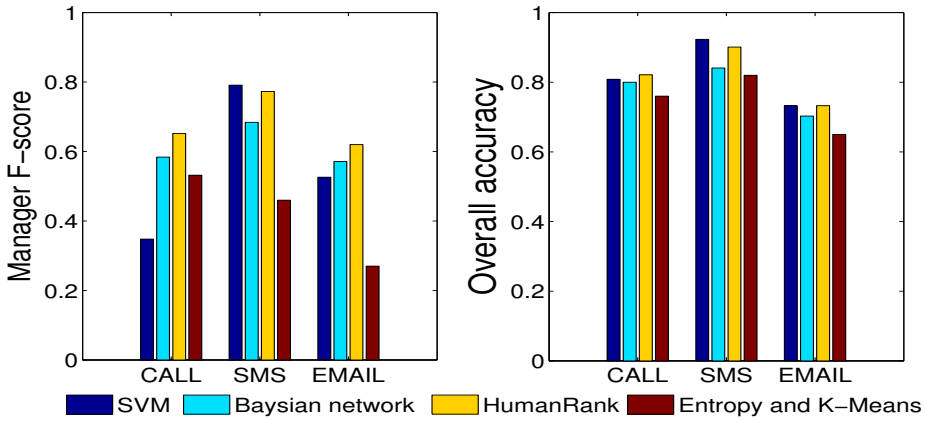


Figure 3.7: Accuracy and Manager's F-score in hierarchy detection using HumanRank and K-means (k=2). We also show measure results of two supervised learning methods, and one unsupervised method using the ranking from [1] and K-means (k=2).

degree of its neighbors, and the average distance to all other nodes in the graph. We then use 10-fold cross validation to generate the results in the chapter. That is, we used 90% of the data for training and 10% for testing. We repeat this several times until all the instances are included in the testing and training at least once. For the unsupervised methods, we repeat the same process as in our classifier, but we rank nodes using the method in [1].

We applied our classification results on all three CINs. In Figure 3.7 (a) and (b), we show the accuracy of our approach compared to different methods. As we see, our method gives consistently good results. Even though in some data sets another method can be better, we found the performance of these approaches to vary significantly across different networks. Especially, HumanRank outperforms other methods in each CIN in terms of manager identification, which is an important advantage of HumanRank, because inferring key members in an organization is more meaningful than inferring ordinary members.

3.7 Summary and Conclusions

With our work we aim to give an in-depth study of the communication patterns formed by employees interacting in two modern enterprise organizations. We call these graphs Communication Interaction Networks, or CINs. Towards this end, we use data from two different organizations; the Enron email dataset, and data from the exchange of short messages and phone-call from a communication provider. Our data are unique in two ways: (a) we are the first to study CINs using two different organizations, and (b) chapterwe are the first to study two CINs by the same employees and two different communication modes. The key observations from our study are: (a) three CINs share

common features despite of their diverse background; (b) employees that play different roles have distinct communication behavior; (c) CINs carry valuable information about the hierarchical structure of an enterprise, and our HumanRank method can correctly identify the job titles for above 70% of their employees.

We hope that our observations and proposed algorithms can find their way to other applications that use interaction networks. For example, it will be very interesting to see if we can identify leaders in criminal or terrorist organizations using only the observed interactions.

Chapter 4

Inference of the demographic properties of users

4.1 Introduction

Homophily, the social phenomenon that individuals tend to bond with similar others, has been discovered in social networks [33]. Past work on homophily is mostly based on surveys, known communities, and more recently online social networks (OSNs) [34]. There has been recent work on OSNs that attempts to infer users' demographic attributes using information from their peer groups [35, 36, 37]. The increasing penetration and usage of cell phones and the availability of their call logs bring another opportunity to study and potentially exploit homophily.

The goal of this chapter is to see whether homophily exists in call networks and if so, to what degree we can infer a cellphochapterne user's demographic properties by knowing

the demographic information of people (s)he talks to. Specifically, we attempt to address a series of questions: 1) Does homophily exist in call networks? 2) What features (related to communication, graph structure, etc.) indicate stronger homophily between two cellphone users? 3) How to exploit homophily so that we can infer the demographic properties of certain subscribers using demographic properties of their social(call) friends? We focus on three types of demographic information: a) home location, b) age group, and c) income level.

Our works is largely motivated and similar to recent studies on Online Social Networks (OSNs) that exploit information from online friends to infer users' demographic profiles. Backstrom *et al.* predict the locations of Facebook users using locations of their friends [36]. Facebook users' ages are estimated by Dey *et al.* in [35]. Both age and country of residence are inferred by MacKinnon and Warren in [37]. It is worth mentioning that our work is not a duplication of these studies. Little is known about whether or not homophily exists in call social networks, or call networks; and if it exists, how to exploit homophily given the communication features and graph-structural features of the call networks.

Call Detail Records (CDRs) have attracted attention from both academia and industry in studying people's social, mobility, and communication behaviors and there has been some recent work on inferring location and/or age/income information of subscribers using CDRs. Some work studies mobility patterns to infer the important places of cellphone users including home and work places [38, 39, 40]. Some other studies attempt to link communication behaviors with demographic profiles and then utilize communication behavior to infer demographics. Mehrotra *et al.* reveal differences between men and women in phone

usage [41]. Frias-Martinez *et al.* predict genders using communication and structural features in [42] and infer socioeconomic levels in [43] for subscribers in a Latin American country. Blumenstock *et al.* have done a series of studies using CDRs from Rwanda on the relationship between phone usage and gender and socioeconomic status [44] and propose a method to estimate wealth based on phone usage [45]. Most of these studies lack ground truth information and have to either estimate it with another method (surveys for example, was conducted to estimate wealth levels), or using aggregated statistics (such as census). Our work differs from these studies in that we use homophily to predict demographics instead of communication features directly, and we conduct individual instead of aggregated inference.

We conduct an extensive study on a nationwide cellular network with over twenty million subscribers observed over one month. We construct three call graphs according to different time slices: the weekday graph, the night-and-weekend graph and the complete graph.Our contribution can be summarized as follows:

- We observe the existence of homophily in terms of home location, age group and income level on call networks. For example, 80% of a subscriber's (call) friends reside within 100 km from him/her. About 60% of wealthy subscribers' friends are also wealthy. However, we notice that around 20% users do not have a friend with the same age group or income level.

- We discover that people interact with friends sharing different demographic attributes at different times. For example, age homophily has a stronger appearance on the weekday graph than on the night-and-weekend graph or on the complete graph.

- We develop effective algorithms to infer home location, age group, and income level. Home locations can be predicted within 20 km radius with 80% probability and we achieve accuracies of 78% and 72% for age and incoming prediction, respectively.

The work is a first step towards understanding how much information can be extracted from a communication network. This could be useful in anti-terrorism and police work, to more benign target-marketing efforts. Due to space limitation, we will elaborate and showcase such applications in future work.

The rest of the chapter is organized as follows. We describe our dataset in Section 5.3. In section 5.4, we present evidence of the existence of homophily on call graphs and introduce its correlations with communication and graph structure properties. In section 4.4, we develop algorithms to infer demographics and measure the accuracy of our proposed algorithms. We conclude in section 5.6.

4.2 Problem definition and data introduction

We use anonymized Call Data Records (CDRs) from a cellular service provider collected during a month in 2010. The dataset consists of over 27 billion call records and 93 million callers out of which 25 million are the provider's subscribers. A record in that dataset consists of two anonymized phone numbers, call direction, start time, end time and the location of the base station that carried the call. For a set of anonymized subscriber numbers (as opposed to the whole set of 25 million for certain reasons), we have associated age group and income level for the subscriber. Home location of the subscriber is inferred from the CDRs as location of the base station that carried the most night-time calls of the

subscriber.

Graph formation. We use CDRs to construct a call graph where its nodes are callers and edges represent bi-directional communication between pairs, meaning both users must call each other. We only consider bi-directional communication, because this reciprocity has been shown to be a good indication of a strong relationship [46].

We have two types of users: the operator's callers, which we refer to as **subscribers**, and callers of other networks. The term user or node describes all users irrespective of their mobile carrier. Note that for non-subscribers we do not have information of their demographics, thus we cannot evaluate the accuracy of our inference. Therefore, we only study node properties of subscribers and edge features of subscriber-subscriber pairs. The **degree** of a subscriber is the total number of communicating friends, including subscribers and other users. By contrast, the term **subscriber-neighborhood** or simply

Table 4.1: Basic statistics of data sets

	weekday	night-weekend	complete
Num. of nodes	93 million	89 million	93 million
Num. of edges	146 million	140 million	213 million
Num. of subscribers	25 million	25 million	25 million
Ave. Num. of calls	10.9	10.8	14.6
Ave. Average call duration	258 secs	296 secs	275 secs
Ave. Num. of Common nodes	0.65	0.81	1.11
Ave. degree difference	43	31	40

neighborhood of a subscriber refers only to other subscribers among all communicating pairs.

Time slices. We use three types of graphs based on the time of data collection, which we refer to as time slice: (a) weekdays, (b) night-weekend, and (c) all interactions. The reason for using time slices is that earlier work has shown that human interactions vary along these different time slices [47]. The weekday slice includes activity between 9:00 a.m. to 6:00 p.m. on weekdays. The night-weekend slice includes records from 9:00 p.m. to 8:00 a.m. on weekdays and all hours on weekends. Based on these time slices, we construct and use three call graphs: **a weekday graph**, **a night-weekend graph**, and the **the complete graph** which includes all recorder communication interactions.

Demographic properties of a subscriber. We select the following demographic properties: age group, income level and home location. We have four age groups: 18-30, 31-45, 46-65 and 66+, which consist of 14.94% , 40.31% , 33.62% and 11.12% of the subscriber population respectively (with known information). We have five income levels: wealthy, affluent, middle, low middle and low, which represent 12.09% , 34.94%, 13.41%, 19.48% and 20.61% of the subscribers respectively. The home location of a subscriber is unknown, but we estimate it using the location of the base station where his/her calls are transmitted most frequently on weekday nights (from 9:00 p.m. to 8:00 a.m.), which is a (longitude, latitude) pair denoted as h. Note that this heuristic for finding home location has been shown to work quite well for call networks [38].

Features of an edge. We enrich the concept of a subscriber-subscriber edge using four features: the number of calls, average call duration, degree difference and the

number of common friends. Degree difference means the absolute difference between the degrees of the two subscribers. Number of common friends means the number of callers that two subscribers both connect with. The averages of those features over all edges are shown below.

Defining the f_i communication features. If we consider the time slice, we have time-slice-specific versions of the above features. For example, when they are applied on the weekday and night-weekend graphs, we have: number of weekday calls (f_1), number of night-weekend calls (f_2), average weekday call duration (f_3), average night-weekend call duration (f_4), weekday degree difference (f_5), night-weekend degree difference (f_6), number of weekday common friends (f_7) and number of night-weekend common friends (f_8).

There are a large number of other features, such as the percentage of common friends, which we intend to evaluate in future work.

4.3 Observations

We are interested in two inter-related questions: (1) Does a call graph exhibit homophily with respect to home location, age group and income level? (2) Which pair-wise communication features are correlated with homophily?

4.3.1 Quantifying graph homophily

Home location homophily. We quantify the effect of home distance on the formation of an edge in the call graph. We expect to see that people residing closer have a higher chance of calling each other. We define and measure two probabilities. First, $p_1(d)$ is

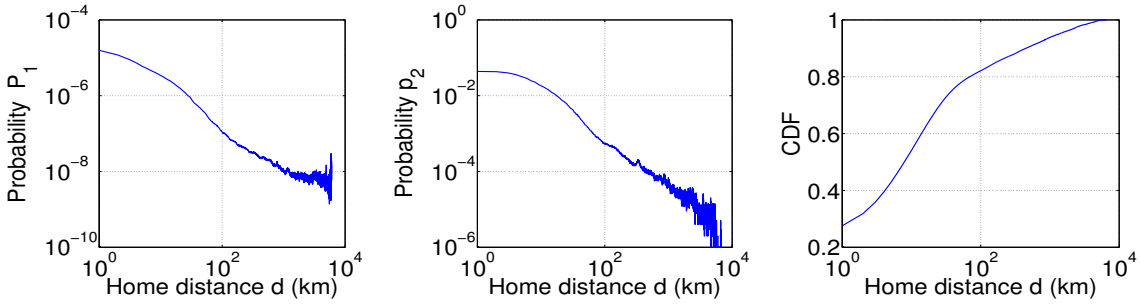


Figure 4.1: The probabilities $p_1(d)$ (LEFT) and $p_2(d)$ (MIDDLE) as a function of d and Cumulative Distribution Function (CDF) of $p_2(d)$ (RIGHT).

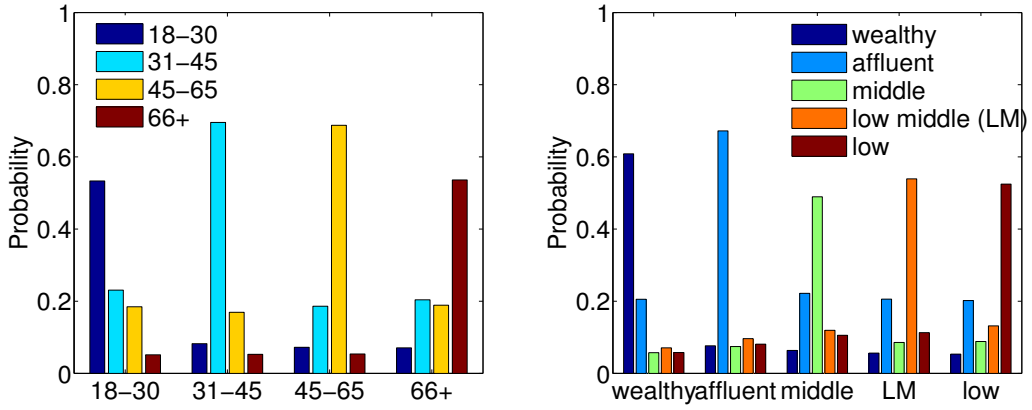


Figure 4.2: People talk to people of same age group and income level with more than 50% probability: Probability of communication among (LEFT) age groups, and (RIGHT) income levels. The x-axis presents a and y-axis presents b.

the probability that two subscribers residing at a home distance d are connected by an edge, that is, $p_1(d) = p(Edge(u_i, u_j)||h_i - h_j| = d)$. Second, given two connected subscribers, $p_2(d)$ is the probability that their home distance is d, that is, $p_2(d) = p(|h_i - h_j| = d|Edge(u_i, u_j))$. u_i and u_j are any two subscribers on a call graph and h_i and h_j are their home locations.

We plot the p_1 and p_2 probabilities on the complete graph in Fig. 4.1. We approximate the tail of the distribution with a power-law distribution $y = ax^b$ for both curves.

Their exponents are about $b = -0.3$ and $b = -1.27$ respectively, suggesting that the formation of an edge has inverse correlation with home distance. The corresponding probability to p_1 on a Facebook friendship graph has been reported to follow a power law with the exponent $b = -1$[36]. Using the cumulative distribution function (CDF) of the $p_2(d)$, for a subscriber, we find that 80% of his/her neighbors reside less than 100km from him/her. The curves from the weekday and night-weekend graphs have similar patterns.

Age homophily. We investigate whether subscribers have high chance of talking to those of the same age. We start with the conditional probability that a subscriber u_i of age a calls the other u_j of age b on the complete graph, that is, $p(AGE(u_j) = b|AGE(u_i) = a)$, which is briefly denoted as $p(b|a)$.

We calculate the probabilities of all the combinations for age groups a and b in Fig. 4.2. We find that **subscribers have more than 50% chance of calling others of the same age among the four age groups.** This finding is an indication of strong homophily regarding age groups. The weekday and night-weekend graphs exhibit present strong homophily and their associated probabilities are very close to those from the complete graph. Interestingly, the weekday graph exhibits slightly stronger homophily, which could be explained that working relationships could favor same-age communication.

Income Homophily: income levels communicate within the same level. With similar analysis as above, we find again a more than 50% probability for people to talk to people of the same income level as shown in Fig. 4.2.

Understanding neighbor dynamics. The existence of homophily provides promise for the possibility of inferring subscriber properties. Here, we take a step closer

and we analyze the neighborhood of each subscriber. The intuitive thought is that: one's age should be the same as the majority ($> 50\%$) of its neighbors. We find that this is true but not for all subscriber neighborhoods. In Fig. 4.3(LEFT), we see that only 60% of subscribers (y-axis) have a neighborhood where the fraction of the same age neighbors is higher than 50% (x-axis). To investigate further, we plot the percentage of subscribers who have k neighbors with similar properties as a function of k in Fig. 4.3 (RIGHT) as a CCDF. Roughly 80% of the subscribers have at least one neighbor (greater than k=0 on x-axis) of the same age and income.

This suggest an alternative approach for our inference compared to a "majority vote": is there a way to identify the neighbor that is most similar to the subscriber in question among all neighbors? To answer this, we would have to look at other communication features between the communicating neighbors that can indicate this "most similar neighbor". Thus, we continue further to study the pair-wise homophily and its correlation with communication features which we defined in section 5.3.

4.3.2 Communication features as indicators of homophily

Our goal here is to find communication features between a pair of subscribers that would suggest a higher similarity of demographic properties. For example, if I talk to you for half an hour per call at night, does this imply we live closer or further? If we have 5 common friends, are we more likely to be in same age group or income level? To identify such correlations, we use linear regression to identify significant features for demographic homophily over all possible pairs on the complete graph. We establish a linear regression

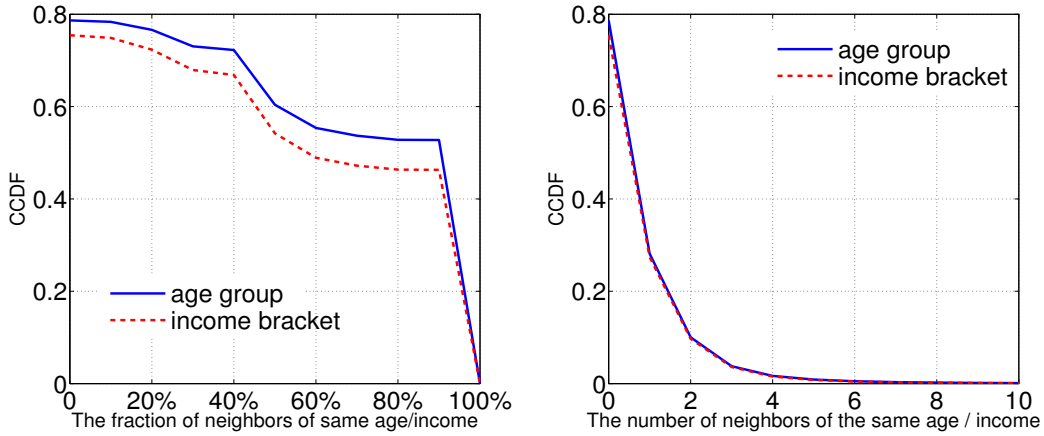


Figure 4.3: Complementary cumulative distribution functions (CCDF) of (LEFT) the percentage, and (RIGHT) the absolute number of neighbors of the same age/income.

model $\delta(u_i, u_j) = \sum_{x=1}^{8} \beta_x f_x(u_i, u_j)$, where $\delta(u_i, u_j)$ is the distance of two homes$|h_i - h_j|$ for home homophily. Recall that f_x are the features that we defined in section 5.3. For age/income, $\delta(u_i, u_j)$ is 1 if their age/income is same, otherwise, it is 0.

Home location homophily. With the regression, both average weekday and weekend-night call durations f_3 and f_4 and the number of weekday and night-weekend common friends f_7 and f_8 are proved to be more important than others. The four show statistical significance with p-value < 0.05 under t-test.

The regression coefficients reveal interesting relationships between calling features and home distance of a communicating pair. First, the regression coefficients of average weekday and night-weekend call durations β_3 and β_4 are 0.075 and 0.069 in the linear model. This suggests that the average call duration is positively correlated with home distance. Second, the coefficient of number of weekday β_7 = -10.018 and night-weekend common friends β_8 = -30.933 show that the number of common friends in weekday and

night-weekend time-slices is negative correlated with home distance. The average distance between the homes of a pair with 5 or more common friends is 50km, while the distance for a pair with 15 or more common friends decreases to 10km.

Age homophily. Regarding age, all the eight features show statistical significance with p-value < 0.05 except the number of night-weekend calls f_2 and number of night-weekend common friends f_8.

The regression coefficients reveal interesting relationships between calling features and age group. First, the coefficient associated with number of weekday calls β_1 is 0.0245, which suggests that people make more weekday calls to those of the same age group. If two subscribers have more than 20 weekday calls, the probability is as high as 80%. Second, the coefficients associated with weekday and night-weekend call duration β_3 and β_4 are around -0.00025, indicating people on average make shorter-duration calls to those of the same age group. For example, calls with average call duration less than 10 minutes on weekends have 50% chance of happening between two subscribers in the same age group. Finally, the coefficient associated with the number of weekday common friends β_7 is 0.30, indicating that two subscribers with more weekday common friends show higher probability of being in the same age group. When two subscribers have more than one weekday or night-weekend common friends, the probability of having the same age is as high as 75%.

Income homophily. We find that all the features show statistical significance with p-value < 0.05 except the number of night-weekend calls f_2. The correlations between features and income level are similar to those between the same features for age homophily, and in fact the regression coefficients have similar values. Because of that, and due to space

limitations, we do not re-state them.

4.4 Inference of home location, age and income

We attempt to infer demographic properties using three different algorithms, Maximum Likelihood (ML), majority vote (MAJ), and RANK, which attempts to find the most similar neighbor. Specifically, RANK orders the neighbors according to "strength of similarity" and labels the subscriber in question with the most likely to be similar neighbor. The algorithms are further explained below. We also consider all three graphs, weekday, night-weeked and complete. For ease of discussion, we number each inference attempt with an ID using a different different graph and infer: (a) location in attempts 1 to 6, (b) age group in 7 to 12, and (c) income level in attempts 13 to 18 in Table 4.2.

We randomly select 1 million subscribers for experiments. For algorithms ML and MAJ, we use all the data for testing. For algorithm RANK, we use 5-fold cross-validation. Each round uses 80% for training and 20% for testing.

Inferring home location. The question here is to assess how accurately we can predict the location of a subscriber's home based on the home locations of his/her neighbors. We select two algorithms to test the accuracy of home location prediction: 1) maximum likelihood (ML)[36] , and 2) our RANK approach.

Given the home locations of the neighbors NS, ML calculates the probability of each home h' by $PX(h') = \prod_{u_k \in NS} p_1(|h' - h_k|) \prod_{u_k \notin NS} (1 - p_1(|h' - h_k|))$, where p_1 is shown in Fig. 4.1 and picks the one with the largest PX as the answer.

RANK uses linear regression as explained in section 5.4 with carefully selected

Table 4.2: The results of inferring demographic properties

ID	call graph G	Algorithm	accuracy	ID	call graph G	Algorithm	accuracy
1	weekday	ML	5.1km	4	weekday	RANK	5.2 km
2	night-weekend	ML	5.1km	5	night-weekend	RANK	5.1 km
3	complete	ML	4.0km	6	complete	RANK	4.0 km
7	weekday	MAJ	78%	10	weekday	RANK	77%
8	night-weekend	MAJ	65%	11	night-weekend	RANK	64%
9	complete	MAJ	72%	12	complete	RANK	71%
13	weekday	MAJ	72%	16	weekday	RANK	71%
14	night-weekend	MAJ	58%	17	night-weekend	RANK	57%
15	complete	MAJ	67%	18	complete	RANK	66%

features. It uses the features of each target-neighbor pair to estimate the distance between target's and neighbor's homes, selects the nearest neighbor and returns its home location as the answer. We use f_3 and f_7 for the weekday graph, f_4 and f_8 for the night-weekend graph and all for the complete graph. The reported accuracy of prediction is defined to be the median from distances between the inferred home location l and the actual one h over all subscribers.

Using ML, we find that the complete graph is better at home inference than the weekday and night-weekend graphs. The accuracy of experiment 1 and 2 is 5.0 km and the accuracy of experiment 3 is 4.3 km. To highlight the difference, we also calculate the distribution of the distances $|l - h|$ and find that 80% of distances are less than 20km

on the complete graph, but 75% on the weekday graph. Using RANK, we find that its accuracy is similar to that of ML on the complete graph. However, a key advantage is that RANK only needs the home location of the most likely neighbor. This argues that the two algorithms can be equally useful under different scenarios. If we have the call records (ie. communication features), and few of home locations, RANK may be useful, because it use the information of the best neighbor, when known. But, if we only have the graph but without accurate communication features (such as intensity of communication), and many of home locations, ML might be a better choice.

Inferring age group: accuracy as high as 78%. We use two algorithms MAJ and RANK to predict the age group. MAJ identifies the most dominant property among the neighbors of the subscriber, and uses that property to label the subscriber in question. The algorithm RANK uses the features of each target-neighbor pair to estimate the likelihood of their similarity, selects the most likely similar neighbor and returns its label as the answer, and we use f_1, f_3, f_5 and f_7 on weekday graph, f_4 and f_6 on the night-weekend graph and all of them on the complete graph. The accuracy is defined as the ratio of the number of subscribers with correctly inferred age over the total number.

We find that the accuracy of RANK is comparable to that of MAJ, within 1% difference. Using MAJ, we find that the weekday graph gives better results than the night-weekend graph: comparing 78% (ID= 7) with 65% (ID=8) in Table 4.2. We also attempt to understand the source of misclassification. For this, we study the confusion matrix for inference on the complete graph (ID=9) in Table 4.3. We find that 31-45 is the age group that is most likely misclassified.

Table 4.3: Confusion matrix of inferences IDs = 9 and 15. ACT=Actual, PRE=Predicted

ACT \ *PRE*	18-30	31-45	46-65	66+		wealthy	affluent	middle	low mid	low
18 - 30	72K	37K	25K	5K	wealthy	89K	29K	5K	7K	6K
31 - 45	15K	325K	38K	9K	affluent	10K	325K	11K	18K	17K
46 - 65	12K	69K	274K	8K	middle	3K	35K	67K	13K	13K
66+	4K	23K	17K	62K	low mid	4K	46K	7K	110K	21K
					low	3K	38K	6K	13K	93K

Going a step further, we analyze whether subscribers with different subscriber-degree are handled differently by the two algorithms. Using the complete graph, we find that RANK outperforms MAJ on low-degree subscribers (d<7) but falls behind on high-degree subscribers (d>7). This comparison shows that two methods can be used synergistically: We can utilize RANK for subscribers with low degree, and we can employ MAJ for higher degree nodes.

Inferring income level. We find several similarities between income and age inference. The weekday graph is again better at income inference than the other two. RANK is more suitable for low-degree subscribers (< 10) and the MAJ algorithm is suitable for high-degree ones (> 10). We use all the features except f_2.

We are also interested in whether the use of a phone can indicate your age group or income level. We use the dataset where uses of phones of a subscriber are recorded, which will be introduced in the next chapter, to extract features to infer the two kinds of demographics. The result is that it is hard to rely on the current and used phones to infer

your age group or income level. The accuracy is around 50%.

Moreover, we also study the usefulness of features of communications extracted from CDRs for predicting age group and income level. They are total call frequency of a subscriber, his/her average talking duration, total number of short messages (s)he sends and his/her total bytes of consumed data. The accuracy is still around 50%.

4.5 Conclusions

In this chapter, we check the existence of demographic homophily on a call graph: home location, age group and income level and explore how to utilize the homophily to infer those demographics. First, we create three kinds of call graphs according to different time slices: the weekday, night-weekend and complete graphs. Second, we find all three call graphs present demographic homophily. Third, we quantify the effect of pairwise communication and topology properties on the homophily. Finally, we measure the accuracy of demographics prediction on three call graphs by using homophily and correlated features.

Chapter 5

Patterns of Phone Switching and Inference

5.1 Introduction

In recent years, the rapid development of the mobile application market and mobile commerce has driven smartphone adoption. Estimates state that there are now more than 1.5 billion smartphones in the world and that 81% of U.S. mobile subscribers have 3G/4G subscriptions. Another study shows that the average life of a smartphone is 18 months. People change phones so frequently that it it critical for telecom operators to study the evolution of smartphones and to keep the pace with the era.

Predicting phone preference is an important topic in the study of the evolution of phone popularity. Previous efforts have studied the correlation between phone functionality and personal preference[48][49]. However, these early efforts have two limitations. First,

they focus on the evolution of cellphones as smartphones have been replacing cellphones. Thus, observations provided by those previous works are out-of-date; (2) they depend on surveys to ask people about their preferences over limited choices but are unable to know actual phones to which they switch .

In this chapter, our goal is to predict future phone preferences accurately. We are interested in the following problem *given the properties of subscribers, such as demographics and previous phone choice, how can we infer preferences for their next phones?* The input of the problem is a group of k subscribers with intents of switching phones and their properties, the output is the numbers of different types of phones for these subscribers. In order to address this problem, we decompose the problem into two sub-questions: 1) what are the trends in switching phones, and are they correlated to user properties? 2) Based on these trends, how can we estimate the demand for new smartphones?

Our work is largely motivated by the need for telecom operators to anticipate phone demands. Currently, it is difficult to predict what phone a subscriber may use next and to determine on what to base that prediction. Inaccurately predicting demand for phones leads to financial loss, especially when ordering expensive smartphones, which cost more than ordinary cellphones.

We conduct an extensive study on a nationwide cellular network dataset with 3 millions subscribers. For each subscriber, we know his/her age groups, income levels, previously used phones and the phone preference of his/her friends (1) who are on the same bill plan and (2) whom they frequently communicate with. In contrast to previous work, we know the actual phones those subscribers switch to, which is more accurate than relying

on survey data. Our contributions can be summarized as follows:

(1) Demographics and information about the previous phone is strongly correlated with the next phone preference. For example, 90% of subscribers who are both rich and young prefer to use smartphones while 65% of subscribers who are both poor and old like to use them.

(2) Social influence affects phone switching for subscribers on the business plan. For example, 75% of them use the same phone as others on the same plan while the corresponding fraction on non-business plans is only 50%. This shows that the phone preference is more predictable for business persons.

(3) Given the properties of a group of subscribers, we build classifiers using Bayesian Network to infer their choices. Compared with previous methods, our approach reduces the prediction errors by 1/3 and the related costs by half.

The rest of the chapter is organized as follows. We describe our dataset in Section 5.3. In Section 5.4, we present correlations between the phone preference and demographic properties, previous phone choice and the effect of influences from friends respectively. In Section 5.5, we develop algorithms to predict demand for phones and measure the accuracy of our proposed algorithms. We conclude in Section 5.6.

5.2 Related Work

Predicting the purchase intent of a phone has attracted much attention in the fields of social, management and computer science. People found functionality and mobile application are important features influencing the purchase of a cellphone[48]. Safiek Mokhlis et

al. [49] found that innovative features, such as built-in camera, significantly affect personal choice in the smartphone. Yuri Park et al. [50] showed the strong correlation between phone adoption and mobile applications in Korea. Matti Haverila [51] revealed business functionality has a significant correlation with repurchase intent. All the literatures considered more functional features than ours, such as input style, touch-screen and keyboard. These data is not available in our dataset. Our work considers more on the properties of subscriber. Moreover, we have features, for example, the use of the previous phone, from the phone-switching history of subscribers to study correlation between uses of previous and current phones.

Feature selection plays a key role to the accuracy of the prediction. Previous efforts listed below inspire us to use demographics and social network-related features. Okazaki et al. [52] revealed the effects of demographic characteristics, including the age, gender, marital status, and occupation, on the adoption of mobile contents. June Lu [53] found that the social influences are potential but not direct determinants of the adoption of wireless Internet service.

The work of inferring the purchase intent of a phone is on the beginning stage and few works is known on the prediction of purchase of a smartphone. Erin Strauts [54] made the prediction of cell phone and landline from a social survey. Derek Strauts [55] mined purchase history to predict adoption of mobile computing, which is most similar to our work in terms of dataset. In two works, logistic regression was used as the prediction model. However, Sergiu Nedevschi [56] also proposed to use Bayesian Network as the model. In this chapter, we use the two models for the demand prediction and compare their

performance.

5.3 Datasets and features

We show the overview of our dataset, define features from the dataset and discuss its possible limitation to the result of this chapter.

5.3.1 Dataset introduction

Dataset of phone-switching history Our dataset consists of over 3 million subscribers who switch phones from Januray to June 2012. We know demographic properties of subscribers from the third party, including age group, income level and the life of his/her used phone and what phone they are using. Of those subscribers, 90% have only one used phone and 10% have more than one used phone. For simplicity, we only study one used phone, so for the 10% we only consider their latest used phones as the used phone. Table 5.1 shows a concrete example where the subscriber 49 is young, his income level is Middle and he used Samsung Galaxy III from 07/01/2011 to 01/31/2012 and switched to iPhone 4s on 02/01/2012.

There are more than 200 types of phones in the dataset. We cluster these phones in the level of operating system (OS) into six groups: iOS, Android, Blackberry, Palm, Windows and Cellphone. The cellphone group mainly includes non-smartphones installed with operating systems except the previous five. Note that in the rest of the chapter, when we mention the term "phone", we do not mean a particular phone but the name of a group of phones. In Table 5.1, Phone 1 is an Android phone because Samsung Galaxy III is

installed with Android OS and Phone 2 is an iOS phone.

Dataset of CDRs and bill plan. First, we collect Call Data Records (CDRs) generated by the 3 million subscribers in 03/2012, which is named as CDR-Mar. This dataset contains all their communication records. In a record there are a caller, a callee, the direction of calling, etc. Using these records, we construct a call graph where nodes are subscribers and edges represent two subscribers calling each other at least once. In the chapter, **a call friend** means a 1-degree neighbor on the call graph. Second, we also get the information of bill plans of the 3 million subscribers in 06/2012, which is named as BILL-Jun. This dataset contains the bill plan to which a subscriber belongs so we know that "who is on the same bill plan with whom". Here, **a plan friend** means a subscriber who is on the same bill plan.

Table 5.1: An example of the phone-switching history of a subscriber

Field	Value	Field	Value
ID	49	Age group	Young
Income level	Middle		
Phone 1	Samsung Galaxy III	OS 1	Android
Start day	07/01/2011	End day	01/31/2012
Phone 2	iPhone 4s	OS 2	iOS
Start day	02/01/2012	End day	none

5.3.2 Definition of features

The goal of this chapter is to infer the phone preference of a subscriber. Here, we consider the following features that may be useful for the inference.

Demographic properties. For a subscriber, we have two types of demographic properties: age group and income level. We have four age groups: young, middle, middle old and old aged 18-30, 31-45, 46-65 and >66, respectively. Their fractions are 17.8%, 43.8%, 29.5% and 8.6% of the whole population. We also have three income levels: low, middle and high. Their percentages are 33.1%, 33.3% and 33.6%, respectively.

Previous phone preference. There are two related features: the used phone denoted as PH_{prev}, and the length of its life denoted as Len. The fractions of six types of PH_{prev} are Android: 42.5%, Cell: 39.7%, Blackberry: 14.8%, iOS: 2.1%, Palm 0.4% and Windows 0.4%. Len is the number of months from the start day to the end day for the use of PH_{prev}. Its Probability Density Distribution (PDF) is shown in Figure 5.2. In the Table 5.1, PH_{prev} is "Android" and its Len is 7 months. With these two features, we can study the correlations between previous and current phones.

Features of social networks. We are also interested in the social influence from friends on the phone preference. There are two kinds of friends: call friends and plan friends. With the definitions of plan friends, we can define **a plan neighborhood** which is a set containing all the plan friends. A plan phone PH_{plan} is the phone that is used most by those friends just before the switching date. For example, the subscriber 49 switches phone on 02/01/2012 and his PH_{plan} is the most popular phone used by his plan friend on 01/31/2012. We use the similar way to define a call neighborhood based on call friends and

define PH_{call} based on a call neighborhood.

In terms of the number of subscribers on a plan, we can divide plans into two categories: business plans containing more than 8 subscirbers and non-business plans having equal or less than 8 subscribers. Subscribers in the business and non-business plans are called **business subscribers** and **non-business subscribers**, respectively. As we see in the Section 5.4, the business subscribers have different switching patterns from non-business ones.

Features of communications. We also study the usefulness of features of communications extracted from CDR-Mar for predicting the next phone. They are total call frequency of a subscriber, his/her average talking duration, total number of short messages (s)he sends and his/her total bytes of consumed data. However, they are not as significant as those listed above under the statistical testing so we won't discuss them further in the rest of the chapter.

Limitation. The datset provider began to release iPhones from October 2011, so it is too late for us to collect enough information related to iPhone-switching patterns. Only 2.1% of subscribers had iPhone as the used phone while for most iPhone users, we cannot know their next choices. The relatively small samples may cause biased results.

5.4 Observations and trends

First, we observe some fundemental statistics w.r.t when and what people switch. Second, we discuss basic trends in phone popularity. For example, which phone brand is more "sticky"? Third, we discuss how social influence affects phone switching.

Table 5.2: Definitions of symbols

Symbos	Definition
Age	Age group
Income	Income level
PH_{prev}	used phone
$Dura$	the length of the life of PH_{prev}
PH_{accout}	phone used most by subscribers on the same plan
PH_{call}	phone used most by 1-hop neighbors on the call graph
PH_{curr}	current phone

Can your demographics tell your phone preference? We explore the correlation between the demographics of a subscriber and his/her phone preference. We calculate the conditional probability $Prob(PH_{curr}|Age)$ that a subscriber uses a phone PH_{curr} given his/her age group, do the similar experiment for income level and show results in Figure 5.1. First, Android is the top choice for all age groups and income levels. No matter which group a subscriber belongs to, the probability of using Android keeps above 40%. Second, for the young group, their probability of using iPhones is 2 times that of using cellphones. The two numbers are about 30% and 15%. For the rich group, the probability of iPhones is 3 times that of using cellphones and the two numbers change to 35% and 12%. The rich group is more likely to use iPhones than the young group. Third, the probability of using smartphones for subscribers who are both rich and young is 90% while the probability for those who are poor and old is only 65%.

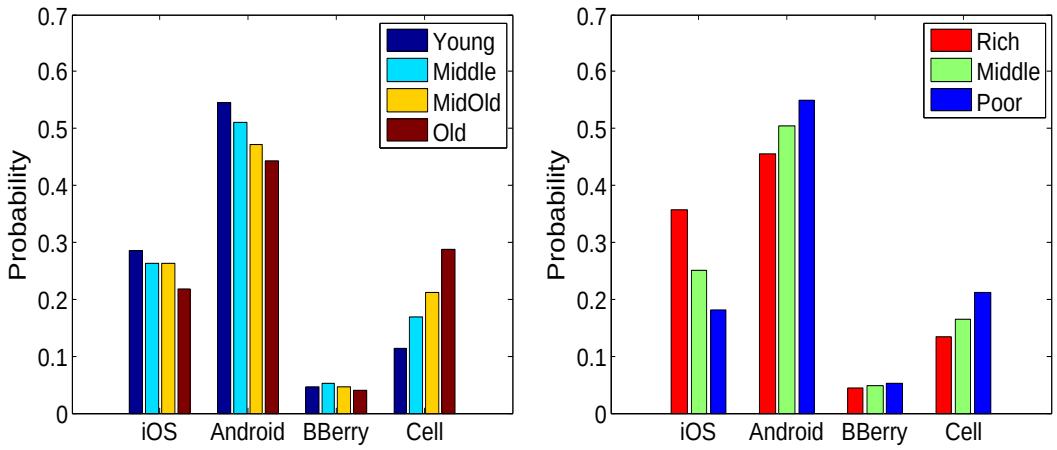


Figure 5.1: The probability that a subscriber uses a phone PH_{curr} given (top) his/her age group and (bottom) income level.

The rich switch phones less frequently than the poor. We plot the probability density distribution of the life of the used phone, which is *Len*, over all subscribers in Figure 5.2. As we see, there are two peaks in the plot: one located at 12 months and the other at 22 months. The second one is within our expectation as people are likely to switch to another phone after they reach the end of a 2-year plan. We are more interested in the appearance of the first peak. Here, we break down the distribution by respectively plotting the versions over subscribers in various age groups and income levels. We find that the plots for various income levels show the reason (See Figure 5.2 (bottom)). For the plot from the poor people, the first peak is more pronounced than the second one but for the curve from the rich people, the second one is as pronounced as the first one. One possible reason could be that the poor subscribers tend to use short-term prepaid phones but the rich ones may like to stably use the service under the contract of a plan.

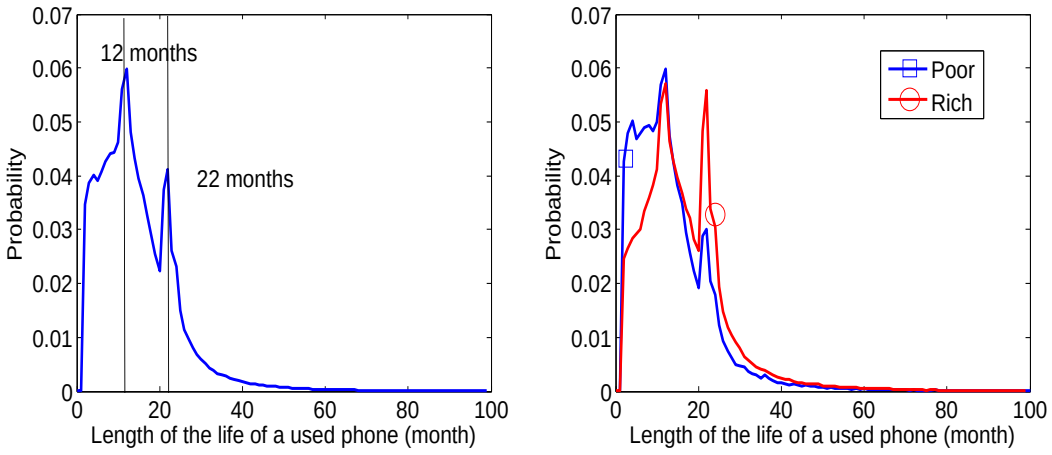


Figure 5.2: Probability Distribution Function of the length Len of the life of a used phone PH_{prev} over (top) all subscribers and (bottom) rich and poor subscribers.

The longer you use a phone, the more you stick to it? We explore the correlation between the length of the life of the used phone Len and the probability you use the same phone by calculating the probability $Prob(PH_{curr} = PH_{prev}|Len = l)$ in Figure 5.3.

The answer is no because the probability doesn't show strictly positve relationships as the Len increases. In Figure 5.3, the plot has two parts divided by a turning point at 2 years. Before the 2-year point, the plot keeps flat, showing that there is no obvious relationship between the two variables. After the point, the plot rises as the duration increases. We find that 80% of subscribers using a phone more than 2 years belong to business subscribers. One possible reason could be that they need to use some specific function of a phone or follow the company policy, so stick to that phone for a long time.

The switching patterns from previous to current phones. Is your previous phone correlated to your next phone? In order to answer this question, we check the

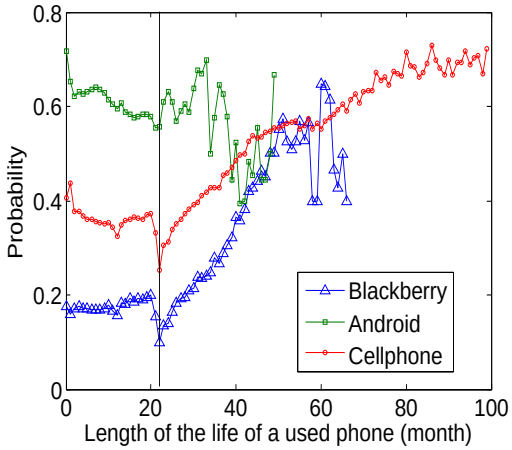


Figure 5.3: The probability that a subscriber uses the same phone PH_{curr} as the used one PH_{prev} given the length of the life of the used phone.

probability that a subscriber use a specific phone given the type of his/her latest used phone, which is denoted as $Prob(PH_{curr}|PH_{prev})$. The result is shown in Table 5.3.

First, Android phones are dominating as the probability $Prob(PH_{curr} = Android|PH_{prev})$ keeps about 45% no matter what the previous phone is. Second, iOS phones are the second dominant phone. In particular, subscribers previously using the Palm and Blackberry phones will equally select iOS and Android as the next phone choice. Third, smartphones are sticky. 90% of smartphone users still choose smartphones. At the same time, 64% cellphone users switch to smartphones.

How does social influence affect phone switching? We are interested in the effect of social influence on the preference of a phone for a subscriber. In other words, we want to see whether one follows the choice of friends when deciding which phone to select. Here we answer this question by checking the influence from plan and call friends as defined in the Section 5.4. Their effects are captured as the probabilities of $PH_{plan} = PH_{curr}$ and

Table 5.3: Probability of switching from the used phone PH_{prev} to the current phone PH_{curr} . The first row and column represent PH_{curr} and PH_{prev}.

	iOS	Android	Blackberry	Windows	Palm	Cell
iOS	0.26	0.53	0.11	0.05	0.01	0.08
Android	0.34	0.57	0.021	0.02	0.03	0.06
Blackberry	0.39	0.42	0.16	0.03	0.001	0.04
Windows	0.34	0.50	0.05	0.04	0.01	0.08
Palm	0.47	0.44	0.03	0.01	0.01	0.04
Cell	0.13	0.48	0.03	0.002	0.01	0.36

$PH_{call} = PH_{curr}$.

We first plot the average probability $Prob(PH_{plan}{=}PH_{curr})$ over subscribers with neighborhood sizes from 2 to 20. As seen in Figure 5.4, there is a strong positive correlation between the probability and the size of a neighborhood when the size is more than 8. This shows that a subscriber tends to follow his/her friends to use the same phone when (s)he is on a business plan. Of all business subscribers, 75% use the same phone as their plan phones. In contrast, on a non-business plan, he/she will not be affected by the influence of his/her plan friends. We see the part of the plot with size ¡ 8 fluctuates and show no rising trend. Only 50% non-business subscribers use the same phone as PH_{plan}.

We also do the similar experiment for the probability Prob($PH_{call}{=}PH_{curr}$). However, we don't observe the correlation pattern as shown previously and the whole plot is flat. We also reconstruct the neighborhood by filtering friends whose call frequency is less than 5 times per month, but we still don't see the similar rising

plot as PH_{plan}. Thus, we won't further consider PH_{call} in the inference models.

5.5 Predicting phone switches

We attempt to infer the next phone to which a subscriber will switch based on his age group, income level, previous phone use, the length of the life of the used phone and the plan phone used by his/her plan friends. The task can be solved as a classification task where the demographics, previous and plan phones are features and the current phone is the outcome. We select two models for the classification: Bayesian Network (Baye) and Logistic Regression (Logi). We use the former because Bayesian Network is able to represent causal relationships among variables. For example, rich people are likely to use smartphones. Moreover, previous work[56] also mentioned to use Bayesian Network to predict the adoption of Information and Communication Technology. We use the latter because logistic regression is widely used in the literature[55][54]. For the two models, there are two steps to do the classification: first, given a group of features of a subscriber i, the model produces a probability p_{ij} that the user i selects the phone j; second, select the phone j with the maximum p_{ij} as the output.

Problem 1: inferring the phone switch of a single subscriber In the whole population, subscribers who switch phones between January and March 2012 are used for the training dataset and the rest are for the testing dataset. From our initial experimental result, it is hard to reach good accuracy for predicting the next phone to which a subscriber will switch no matter which algorithm is employed. The fraction of subscribers with inaccurate phones is as high as 40%. We check the confusion matrix and find that the misclassifications

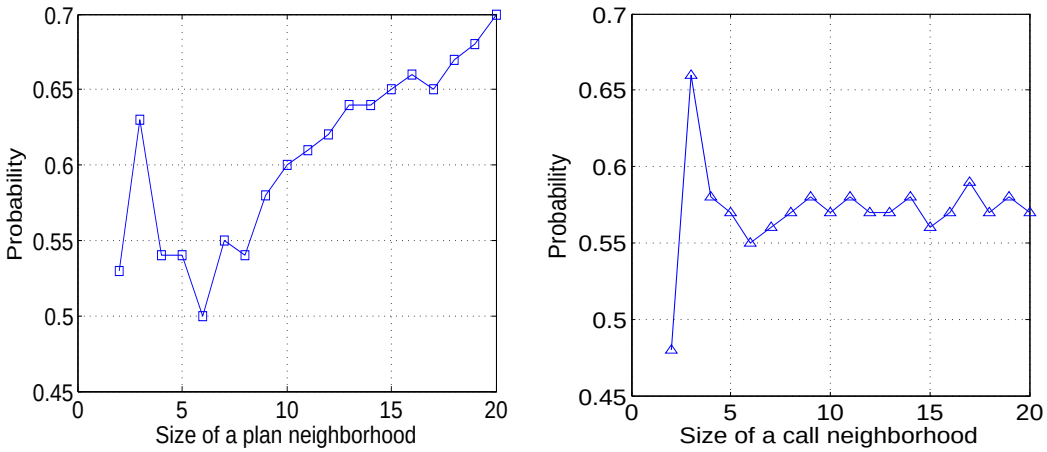


Figure 5.4: Probabilities of (top) PH_{social}=PH_{curr} and (bottom) PH_{plan}=PH_{curr} for the size of a neighborhood from 2 to 20.

from iOS to Android and vice versa contributes to the main part. The possible reason could be that the functions of iOS and Android phones may be so close that it is hard to tell them apart depending on the five features.

Problem 2: inferring phone switches of multiple subscribers We shift to dealing with the problem for multiple subscribers. The reasons are that telecom operators order multiple phones at once for its subscribers, and that the inaccurate prediction due to iOS-to-Android and Android-to-iOS errors is anticipated to cancel each other out. Thus, we attempt to solve the problem for multiple subscribers as follows: *given a group of k subscribers and their properties, how many various types of phones needs to be ordered by inference?* We use the same training and testing datasets as in the Problem 1.

Algorithm We still use bayesian network and logistic regression as the classification models, but treat its output in a different way. Given the probability of phone choices of k subscribers, we add up all the probabilities p_{ij} over the k subscribers for a specific

phone j as the predicted number for the phone, which is denoted as NP_j.

$$NP_j = \sum_{i=1}^{k} p_{ij}$$

where j $\in$ {iOS, Android, Plam, Blackberry, Windows, and cellphone}.

Intuitively, people can make use of the historical trend from previous data. Thus, we employ a method called Histro as the baseline. Histro calculates the fraction of each phone subscribers switch to from the training dataset, multiples it with the number of subscribers from the testing dataset, and obtain the product as the predictedly ordering number for that phone.

Evaluation We evaluate using two metrics for the accuracy of predictions. First, we define a metric called Absolute Difference (AD), which means the difference between the predicted number NP_j for the phone j and its actual number denoted by N_j.

$$AD_j = |N_j - NP_j|, AD = \sum_{j} AD_j$$

Note that AD without j means the summation of ADs over six types of phones.

However, the cost of inaccurately predicting an iPhone is much higher than a cellphone. Table 5.5 lists the cost of over-estimating and under-estimating six kinds of phones and we see that the cost of ordering one more iPhone is 4 times that of an ordinary cellphone. Such the cost is not considered by AD, so we propose the other metric called Absolute Cost Difference (ACD). ACD takes the difference among phone costs into account and use a weight, which is presented by $COST$, to consider the difference.

$$ACD_j = |N_j - NP_j| \times COST_j, ACD = \sum_j ACD_j$$

where $COST_j$ is over-estimated cost $O - COST$ when $N_j > NP_j$. Otherwise, $COST_j$ is under-estimated cost $U - COST$.

The numbers in Table 5.5 are set by domain experts. For the over-estimating cost $O - COST$, it is set according to the price of a phone. If the ordered phone cannot consumed by a subscriber, the operator is not able to return it to a manufacturer. Thus, the cost is roughly proportional to the price of a phone. For the under-estimating cost, it is a constant. If the operator finds no phone for the demand of a subscriber, it can order it from the manufacturer later on and the cost is proportional to the delivering cost, which is set to a fixed value in this chapter.

Bayesian Network approach reduces Absolute Difference (AD) by 30% compared with Histro. The Absolute Difference by Bayesian Network is around 120K while the corresponding number by Histo is around 180K from Figure 5.4, showing that Bayesian Network has better performance than Histro in terms of the aggregated result. In Table 5.4, the ADs of Android and Cellphone by Bayesian Network are about 2 and 1.5 times lower than by Histro. However, AD of iOS by Bayesian Network is about 1.5 times that by Histro.

Adjustment with a cost matrix The second problem wants to minimize the absolute cost difference ACD because the telecom operator prefers saving more money rather than just reducing the absolute difference. For example, mispredicting 2 cellphones might cause less loss than mispredicting 1 iPhone. In order to meet this requirement, we construct

Table 5.4: Result of predictions by Histro, Logistic regression and Bayesian network with/without cost matrix. ↓ and ↑ mean under- and over-estimating.

	iOS	Android	Blackberry	Windows	Palm	Cell
Actual	419,297	909,296	93,368	6,454	6,940	384,942
Histro NP	459,712	959,943	84,973	6,554	4,732	304,383
Histro AD	40, 415 ↑	50,647 ↑	8, 395 ↓	100 ↑	2,208 ↓	80,559 ↓
Logi NP	478,500	895, 812	87,800	2,644	3,312	351229
Logi AD	59,203 ↑	13,484 ↓	5,568 ↓	3,810 ↓	3,628 ↓	33,713 ↓
Baye NP	479,900	895,937	87,583	2,747	3,268	350,862
Baye AD	60,603 ↑	13,359 ↓	5,785 ↓	4,193 ↓	3,186 ↓	34,080 ↓
Baye+Mat NP	370,975	860,894	121,859	4,777	2,934	458,858
Baye+Mat AD	48,322 ↓	48,402 ↓	28,491 ↑	2,163 ↓	3,520 ↓	73,916 ↑

Table 5.5: Under- and over-estimated costs of 6 types of phones

	iOS	Android	Blackberry	Windows	Palm	Cell
$U - COST$	10	10	10	10	10	10
$O - COST$	200	150	100	150	100	50

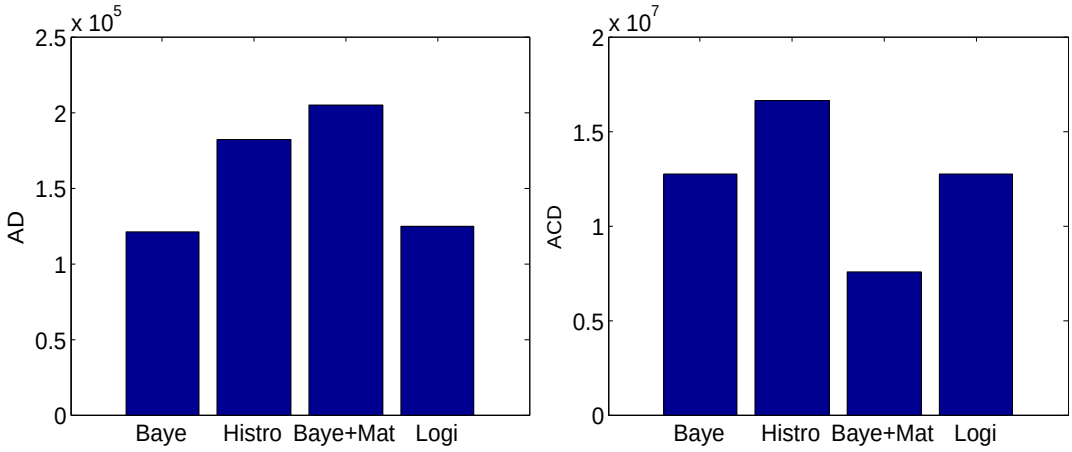


Figure 5.5: The absolute difference (AD) and the absolute cost different (ACD) between groundtruth and the prediction by Histro, Baye, Baye+Mat, and Logi.

a cost matrix with the definition as follows:

$$c_{ij} = O - COST_j / O - COST_i$$

If the $O - COST_j > O - COST_i$, the penalty cost c_{ij} is higher than vice versa because inaccurately ordering a more expensive phone that cannot be consumed by a subscriber means higher financial loss. With this matrix, we anticipate that the inaccurate number of smartphones, especially iPhones, will be reduced.

Bayesian Network method with a cost matrix (Baye+Mat) reduces the absolute cost difference by 50% compared with Histro. We apply the cost matrix on the top of Bayesian Network and show the result in Table 5.4. Comparing with the output without the cost matrix, we get fewer inaccurate iOS and Android phones but more cellphones. However, the cost of inaccurately predicting a cellphone is much lower than a smartphone and the total of absolute cost differences is still

reduced by 1/2 (See Figure 5.5).

Discussion The performance of Histro is typically worst among all the prediction methods regarding both the absolute difference and the absolute cost difference. It inaccurately predicts the most iOS, Android, Blackberry and Cell phones, but we can also observe that Histro is better at predicting the number of Palm and Windows phones. In contrast, Bayesian Network-related algorithms show better performance than Histro for all the phones except the two. The classification model is biased towards the types of the majority because of the imbalance among the numbers of different phones in the training process. As a result, it tends to give a higher probability to iOS than to Windows. Compared with Bayesian Network, Logistic Regression is also a competitive candidate for building the model in terms of both metrics, but the time of building its model is much longer than that of training a Bayesian Network. We use a sample of 20,000 records for training, the time of Bayesian Network provided by Weka is only 2 seconds while that of Logistic Regression is 51 seconds.

5.6 Summary and Conclusions

In this chapter, we show a case study about phone preference on a dataset provided by a telecom operator. Compared with previous work, our work put emphasis on: (1) study the patterns of smartphone switches of subscribers; (2) how the patterns correlate with their demographic properties and social influence from their friends; (3) use the real-world data to design a Bayesian Network based algorithm to infer phones multiple subscribers switch to. The ultimate goal of our work is help telecom operators forecast the phone inventory

and effectively reduce the loss due to the misprediction.

Chapter 6

Conclusion

The dissertation is a systematic effort to model and study the measurement of multimodal communication networks as well as the inference of subscriber demographics by the communication-related features emerging from those networks. First, I present various measurements of how multimodal communications are used among people of different age groups and income levels both in a city and in a company. Second, I propose three groups of algorithms to predict their job titles, home locations/age groups/income levels, and phone preferences, respectively.

With the experimental results in each chapter, I conclude with two suggestions. First, the proper use of time slices and communication modes boosts the inference accuracy significantly. In the dissertation, we consider three kinds of time slices: weekday, weekend, and holiday, and two communication modes: calling and texting. With various times slices and modes, I am able to construct graphs with differing potential for knowledge discovery. For user-role extraction, the texting graph is generally superior to the call graph because

managers have higher degree than ordinary employees in the texting graph than in the calling graph. As a result, both are more distinguishable from one another via the texting graph. For demographics inference, the weekday graph is superior to the night-weekend graph at both inference of age group and income level. The reason is that calling on weekday mornings is more likely to occur among colleagues than on nights or weekends, and colleagues are more likely to be in the same age group or of the same income level. Thus, I suggest identifying what kind of information people want to discover before the construction of a graph and using proper time slices and communication modes to build a graph suitable to the information sought.

Second, relying on the attributes of friends is a reliable method for inferring the demographic properties of a subscriber. According to the results of this dissertation, the accuracy of methods which use the properties of friends is higher than those using other properties. In role extraction, the accuracy of HumanRank reaches 80%, while that of Bayesian Network is only 70%. In demographics inference, the accuracy of MAJ reaches 80%, while that using personal preferences is only 50%. I also show how to utilize the calling features to identify friends with the most reliable information when not all of said friends' information can be obtained. Thus, I suggest considering information of your friends garnered from friends as a potential answer when predicting a certain property of a subscriber.

I also list potential research directions regarding further work. Our call graphs are incomplete because we are not aware of the existence of edges between external subscribers who use other network carriers. Because of this, many graph features, like clustering co-

efficient, cannot be applied in some cases: for example, in the chapter of demographics inference. Assuming that we can turn the incomplete graph into a complete one, I am very interested in how the structure of a subgraph is correlated with age groups of its involved nodes and how the potential new patterns can further increase the inference accuracy. Additionally, I am interested in how to explore the dual-mode graph, where two edges, a call edge and a text edge, may exist between two nodes, from the aspect of methods of measurement and formalizations of related problems.

Bibliography

[1] J. Shetty and J. Adibi, "Discovering important nodes through graph entropy the case of enron email database," in *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, KDD '05, pp. 74–81, 2005.

[2] A. A. Nanavati, R. Singh, D. Chakraborty, K. Dastupta, S. Mukherjea, G. Das, S. Gurumurthy, and A. Joshi, "Analyzing the structure and evolution of massive telecom graphs," *IEEE Transactions on Knowledge and Data Engineering*, vol. 20, no. 5, pp. 703–718, 2008.

[3] L. Akoglu and B. Dalvi, "Structure, tie persistence and event detection in large phone and sms networks," in *Proceedings of the Eighth Workshop on Mining and Learning with Graphs*, MLG '10, (New York, NY, USA), pp. 10–17, ACM, 2010.

[4] J.-P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, D. Lazer, J. K. K. Kaski, and A.-L. Barabási, "Structure and tie strengths in mobile communication networks," *PNAS*, vol. 104, no. 18, pp. 7332–7336, 2007.

[5] J. Leskovec and E. Horvitz, "Planetary-scale views on a large instant-messaging network," in *Proceeding of the 17th international conference on World Wide Web*, WWW '08, (New York, NY, USA), pp. 915–924, ACM, 2008.

[6] J.-P. Onnela, J. Saramaki, J. Hyvonen, G. Szabo, M. A. de Menezes, K. Kaski, A.-L. Barabasi, and J. Kertesz, "Analysis of a large-scale weighted network of one-to-one human communication," *New Journal of Physics*, vol. 9, no. 179, 2007.

[7] L. Akoglu and C. Faloutsos, "Event detection in time series of mobile communication graphs," in *Army Science Conference*, 2010.

[8] H. Zang and J. C. Bolot, "Mining call and mobility data to improve paging efficiency in cellular networks," in *Proceedings of the 13th annual ACM international conference on Mobile computing and networking*, MobiCom '07, pp. 123–134, 2007.

[9] T. Karagiannis, J.-Y. Le Boudec, and M. 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, (New York, NY, USA), pp. 183–194, ACM, 2007.

[10] M. E. J. Newman, "Scientific collaboration networks. i. network construction and fundamental results," *Physical Review E*, vol. 64, p. 016131, 2001.

[11] A. L. Barabasi and R. Albert., "Emergence of scaling in random network," *Science*, vol. 286, no. 5439, pp. 509–512, 1999.

[12] M. Gjoka, M. Kurant, C. T. Butts, and A. Markopoulou, "Walking in facebook: a case study of unbiased sampling of osns," in *Proceedings of the 29th conference on Information communications*, INFOCOM'10, pp. 2498–2506, 2010.

[13] M. Faloutsos, P. Faloutsos, and C. Faloutsos, "On power-law relationships of the internet topology," in *Proceedings of the conference on Applications, technologies, architectures, and protocols for computer communication*, SIGCOMM '99, pp. 251–262, 1999.

[14] H. Jeong, S. Mason, A. L. Barab'asi, and Z. N. Oltvai, "Lethality and centrality in protein networks," *Nature*, vol. 411, pp. 41–42, 2001.

[15] P. Mahadevan, D. Krioukov, K. Fall, and A. Vahdat, "Systematic topology analysis and generation using degree correlations," in *Proceedings of the 2006 conference on Applications, technologies, architectures, and protocols for computer communications*, SIGCOMM '06, pp. 135–146, 2006.

[16] D. Reinhard, *Graph Theory.* third ed., 2006.

[17] D. J. Watts and S. H. Strogatz, "Collective dynamics of small-world networks," *Nature*, vol. 393, pp. 400–442, 1998.

[18] M. E. J. Newman, "Assortative Mixing in Networks," *Physical Review Letters*, vol. 89, pp. 208701+, Oct. 2002.

[19] U. Brandes, "A faster algorithm for betweenness centrality," *Journal of Mathematical Sociology*, vol. 25, pp. 163–177, 2001.

[20] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, "Network motifs: Simple building blocks of complex networks," *Science*, vol. 298, no. 5594, pp. 824–827, 2002.

[21] J. Kleinberg, "Authoritative sources in a hyperlinked environment," *Journal of the ACM*, vol. 46, no. 5, pp. 604–632, 1999.

[22] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: bringing order to the web," *Technical Report, Standford InfoLab*, 1999.

[23] T. Karagiannis and M. Vojnovic, "Behavioral profiles for advanced email features," in *Proceedings of the 18th international conference on World wide web*, WWW '09, pp. 711–720, 2009.

[24] R. Rowe, G. Creamer, S. Hershkop, and S. J. Stolfo, "Automated social hierarchy detection through email network analysis," in *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pp. 109–117, 2007.

[25] M. Gupte, P. Shankar, J. Li, S. Muthukrishnan, and L. Iftode, "Finding hierarchy in directed online social networks," in *Proceedings of the 20th international conference on World wide web*, WWW '11, pp. 557–566, 2011.

[26] A. S. Maiya and T. Y. Berger-Wolf, "Inferring the maximum likelihood hierarchy in social networks," in *Proceedings of the 2009 International Conference on Computational Science and Engineering - Volume 04*, CSE '09, pp. 245–250, 2009.

[27] J. Leskov, J. Kleinberg, and C. Faloutsos, "Graphs over time: densification laws, shrinking diameters and possible explanations," in *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, KDD '05, pp. 177–187, 2005.

[28] Y.-Y. Ahn, S. Han, H. Kwak, S. Moon, and H. Jeong, "Analysis of topological characteristics of huge online social networking services," in *Proceedings of the 16th international conference on World Wide Web*, WWW '07, pp. 835–844, 2007.

[29] J. Abello, P. Pardalos, and M. G. C. Resende, "On maximum clique problems in very large graphs," *DIMACS Series on Discrete Mathematics and Theoretical Computer Science*, vol. 50, pp. 119–130, 1999.

[30] D. Chakrabarti, Y. Zhan, and C. Faloutsos, "R-mat: a recursive model for graph mining," in *SIAM International Conference on Data Mining*, SDM '04, 2004.

[31] Q. Chen, H. Chang, R. Govindan, S. Jamin, S. J. Shenker, and W. Willinger, "The origin of power laws in internet topologies revisited," in *In Proceedings of the 21st Annual Joint Conference of the IEEE Computer and Communications Societies*, INFOCOM '02, pp. 608–617, 2002.

[32] G. Carenini, R. T. Ng, and X. Zhou, "Scalable discovery of hidden emails from large folders," in *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, KDD '05, pp. 544–549, 2005.

[33] M. McPherson, L. S. Lovin, and J. M. Cook, "Birds of a Feather: Homophily in Social Networks," *Annual Review of Sociology*, vol. 27, no. 1, pp. 415–444, 2001.

[34] A. Mislove, B. Viswanath, K. P. Gummadi, and P. Druschel, "You are who you know: inferring user profiles in online social networks," in *WSDM*, pp. 251–260, 2010.

[35] R. Dey, C. Tang, K. Ross, and N. Saxena, "Estimating age privacy leakage in online social networks," in *INFOCOM*, 2012.

[36] L. Backstrom, E. Sun, and C. Marlow, "Find me if you can: improving geographical prediction with social and spatial proximity," in *WWW*, pp. 61–70, 2010.

[37] I. MacKinnon and R. Warren, "Age and geographic inferences of the livejournal social network," in *SNA-ICML*, pp. 176–178, 2007.

[38] S. Isaacman, R. Becker, R. Cáceres, S. Kobourov, M. Martonosi, J. Rowland, and A. Varshavsky, "Identifying important places in people's lives from cellular network data," in *Pervasive*, pp. 133–151, 2011.

[39] F. Calabrese, G. Di Lorenzo, L. Liu, and C. Ratti, "Estimating origin-destination flows using mobile phone location data," *IEEE Pervasive Computing*, vol. 10, no. 4, pp. 36–44, 2011.

[40] V. Frias-Martinez, J. Virsena, A. Rubio, and E. Frias-Martinez, "Towards large scale technology impact analyses: Automatic residential localization from mobile phone-call data," in *ICTD*, 2010.

[41] A. Mehrotra, A. Nguyen, J. Blumenstock, and V. Mohan, "Differences in Phone Use between Men and Women: Quantitative Evidence from Rwanda," in *ICDT*, pp. 297–306, 2012.

[42] V. Frias-Martinez, E. Frias-Martinez, and N. Oliver, "A Gender-centric Analysis of Calling Behavior in a Developing Economy Using Call Detail Records," in *AAAI 2010 Spring Symposia Artificial Intelligence for Development*, 2010.

[43] V. Soto, V. Frias-Martinez, J. Virseda, and E. Frias-Martinez, "Prediction of socioeconomic levels using cell phone records," in *UMAP*, pp. 377–388, 2011.

[44] J. Blumenstock and N. Eagle, "Mobile Divides: Gender, Socioeconomic Status, and Mobile Phone Use in Rwanda," in *ICDT*, 2010.

[45] J. Blumenstock, Y. Shen, and N. Eagle, "A Method for Estimating the Relationship Between Phone Use and Wealth," in *ICDT Workshop*, 2010.

[46] C. A. Hidalgo and C. Rodriguez-Sickert, "The Dynamics of a Mobile Phone Network," *Physica A: Statistical Mechanics and its Applications*, vol. 387, pp. 3017–3024, Feb. 2008.

[47] L. Akoglu and C. Faloutsos, "Event detection in time series of mobile communication graphs," in *ASC*, 2010.

[48] G. Işand G. Büyüközkan, "Using a multi-criteria decision making approach to evaluate mobile phone alternatives," *Comput. Stand. Interfaces*, vol. 29, pp. 265–274, Feb. 2007.

[49] S. Mokhlis and A. Y. Yaakop, "Consumer choice criteria in mobile phone selection: An investigation of malaysian university students," *International Review of Social Sciences and Humanities*, vol. 2, no. 2, pp. 203–212, 2012.

[50] Y. Park, H. Kim, and J. Lee, "An empirical analysis on consumer adoption of mobile phone and mobile content in korea," *Int. J. Mob. Commun.*, vol. 8, pp. 667–688, Sept. 2010.

[51] M. Haverila, "Mobile phone feature preferences, customer satisfaction and repurchase intent among male users," *Australasian Marketing Journal (AMJ)*, vol. 19, no. 4, pp. 238 – 246, 2011.

[52] S. Okazaki, "What do we know about mobile internet adopters? a cluster analysis," *Inf. Manage.*, vol. 43, pp. 127–141, Mar. 2006.

[53] J. Lu, J. E. Yao, and C.-S. Yu, "Personal innovativeness, social influences and adoption of wireless internet services via mobile technology," *The Journal of Strategic Information Systems*, vol. 14, no. 3, pp. 245 – 268, 2005.

[54] E. Strauts, "Prediction of cell phone versus landline use in the general social survey," *The Journal of Undergraduate Research*, vol. 15, 2010.

[55] D. Stanford, "Mining purchase history to predict adoption of mobile computing," in *Intelligent Environments'09*, pp. 482–485, 2009.

[56] S. Nedevschi, J. Sandhu, J. Pal, R. Fonseca, and K. Toyama, "Bayesian networks: A statistical approach for understanding ict adoption," in *IEEE International Conference on Information and Communication Technologies and Development*, 2006.