Inferring Social Relations from Online and Communication Networks

Dissertation submitted for the degree of
Doctor of Natural Sciences

submitted by

Mehwish Nasim

at the

Faculty of Sciences
Department of Computer and Information Science

Date of oral examination: 12.10.2016

1st Reviewer: Prof. Dr. Ulrik Brandes
2nd Reviewer: Prof. Dr. Christophe Prieur (Telecom ParisTech)

dedicated to my father, Mohammad Nasim,
and to my mother, Rukhsana Yasmin,
who offered me every opportunity I ever wanted

Preface

I am glad and excited that my dream of achieving a PhD has come true, *Alhamdulillah*!

It feels like yesterday when Ulrik Brandes invited me for an interview in February 2012. That occasion to interact with him and his wonderful group, aided me to embrace my newly found passion – network analysis. It is with deep pleasure that I extend thanks to my advisor Ulrik Brandes for offering me the opportunity to pursue my PhD under his supervision[1]. Further, perks that this opportunity brought in the form of prospects and liberty to collaborate with other researchers, attendance at conferences, and the independence to work through my ideas, no matter how outlandish they sounded, would go a long way in my future career. The series of unparalleled seminars on social network analysis refined my approach toward addressing research problems and further intrigued my interest in network analysis. I am patently grateful to my advisor for the precious time and resources he vested in me. The convenience of an almost exclusive office space, a contemporary coffee machine, and the chance to brainstorm with the smart *algos*, all turned out to be very helpful in developing this thesis.

Research is no fun without wonderful colleagues – the Christmas presents by Sabine, a birthday lunch at Arlind's, and my first introduction to *Deutsche Kultur* given by David and Uwe in Hamburg, certainly helped me integrate in this multicultural group. I thank all my colleagues for their direct and indirect support toward the completion of this thesis, especially Christine for her extended help whenever I needed it. Her unrelenting support was something I counted on throughout my PhD.

I would like to thank Christophe for his consent to review this thesis, for providing the data, and for the wonderful food in Paris. I am also grateful to Michael and Sven for joining my examination committee. Moreover, I would like to thank all the co-authors, especially Aimal, Raphaël and Usman. I thoroughly enjoyed inter-continental collaborations. Lunch break discussions with Immanuel, and the constructive comments on my work from Barbara, Felix and Habiba were very helpful.

I am grateful to all my relatives and friends who kept me driven by reinforcing my enthusiasm. I would like to say a big Thank You to Viv for being the friend I needed. She was a sister

[1]Financially supported by DEUTSCHE FORSCHUNGSGEMEINSCHAFT(DFG), *Reinhart Koselleck Project*, under grant Br 2158/6-1.

I never had. I benefited a lot from her Statistics knowledge, witty humor, and an unlimited supply of cookies and chocolates. Owais's inspirational attitude towards problems, and Nazo's last minute magical manuscript-proofreads and motivational emails, made my PhD journey smoother.

From giving me an introduction to machine learning methods, to finding time to listen to my research scurries and daily rut, not even a moment passed by when I could not count on you, MJ. Thanks for being an integral part of my life and for the love that you have showered in all these years.

Finally, I would like to thank Abbu and Mamma, for giving me a poised upbringing and for their patience, love and hugs.

Deutsche Zusammenfassung

Soziale Netzwerke bestehen aus einer Anzahl von sozialen Einheiten (Menschen, Akteure, Organisationen etc.), und anderen sozialen Interaktionen von Akteuren. Soziale Netzwerke bestehen nicht nur aus Sammlungen von dyadischen Variablen; Verbindungen in sozialen Netzwerken sind systematisch gemustert und deshalb über die dyadische Ebene hinaus eingebettet.

Die Perspektive der sozialen Netzwerkanalyse stellt eine Reihe von Methoden zur Verfügung um die Strukturen der sozialen Einheiten zu analysieren ebenso wie eine Vielzahl von Theorien um die Muster in sozialen Netzwerken zu erklären. Die Muster zu verstehen, die menschliches Verhalten unterscheidet sind von immenser Wichtigkeit um viele aktuelle Phänomene besser zu verstehen, wie z.B. die Ausbreitung von Innovationen oder Ideen, das Gesundheitswesen, Gruppenbildung und Informationsmanagement um nur einige zu benennen.

Soziale Netzwerke sind zutiefst dynamisch und entwickeln sich mit der Zeit. Längsschnitt Netzwerkdaten, z.B. zu verschiedenen Zeitpunkten gesammelte Daten sind wichtig um einschätzen zu können ob das soziale Umfeld eines Akteurs sein Verhalten beeinflusste oder ob das Verhalten eines Akteurs das Ergebnis einer Änderung der Beziehungen war.

Soziale Netzwerke im Internet sind im letzten Jahrzehnt weltweit zu einem unverzichtbaren Kommunikationsmittel geworden. In dieser Arbeit setzen wir den Schwerpunkt auf die Analyse der sozialen Bindungen und sozialen Interaktionen (mit dem Schwerpunkt in internetbasierten sozialen Netzwerken).

Die Arbeit ist folgendermaßen aufgebaut:

Kapitel 2 In Kapitel 2 erwähnen wir die Präliminarien.

Kapitel 3 In Kapitel 3 analysieren wir die Gruppen (auch Gemeinschaft genannt) in sozialen Netzwerken.

Kapitel 4 In Kapitel 4 analysieren in dieser Arbeit verwendeten Methoden.

Der I Teil der Arbeit ist in 3 Kapitel unterteilt.

Kapitel 5 In Kapitel 5 analysieren wir den Zusammenhang zwischen sozialen Gemeinschaften und Beziehungsmustern wenn Nutzer soziale Netzwerke im Internet verwenden. Wir untersuchen das Beziehungsmuster von Facebook Nutzern und analysieren ob die Änderungen in den veröffentlichten Beiträgen von den vorherigen Antworten auf den Beitrag abhängen oder nicht.

Kapitel 6 In Kapitel 6 analysieren wir die Gruppen (auch soziale Kreise genannt) in sozialen Netzwerken. Wir untersuchen die Zusammensetzung von sich überlagernden sozialen Kreisen eines Egos und den Zusammenhang zwischen den verschiedenen Bestandteilen der sozialen Kreise und den Eigenschaften von Egos.

Kapitel 7 In Kapitel 7 zeigen wir den Einfluss von zusätzlichen Informationen zur Interaktion auf den Rückschluss von Verknüpfungen zwischen Knoten in teilweise verdeckten sozialen online Netzwerken. Wir zeigen, dass Informationen zur Interaktion helfen können, bessere Rückschlüsse auf nicht beobachtete (z.B. fehlende oder verborgene) Beziehungen zu ziehen. Unsere Ergebnisse lassen vermuten, dass in Abwesenheit einer Netzwerkstruktur, Informationen zur Interaktion verwendet werden können stellvertretend für Freundschaftsbeziehungen und somit die Leistung der Vorhersage von Beziehungen verbessern können.

Der II Teil der Arbeit ist in 2 Kapitel unterteilt.

Kapitel 8 In Kapitel 8 analysieren wir Interaktionsverhalten anhand von aufgezeichneten Telefondaten. Wir untersuchen wie viele aktive Kontakte Mobilfunknutzer haben. Wie oft sie angerufen werden. In Bezug auf die Anrufhistorie sind wir an folgendem interessiert: Verteilung der Anrufe, besser gesagt, welcher Prozentsatz der Kommunikation wird mit den Hauptkontakten gepflegt? Und wie oft rufen Menschen die kürzlich erst Angerufenen wiederum an?

Kapitel 9 In Kapitel 9 schlagen wir ein Vorhersagemodell für Telefonanrufe vor, das die zeitlichen Anrufmuster von Nutzern in Betracht zieht.

Contents

1. Introduction

1.1. Motivation: Understanding Social Interaction

Social networks are made up of a set of social entities (people, actors, organizations etc.) and social relations (friendship, kinship, etc.), between those entities. Social relations consists of persistent relations such as *friendship* and instantaneous relations such as *talk to*, *joint participation in an event*, *extend help to*, etc. In the context of this thesis, persistent relations are referred to as social relations and instantaneous relations are referred to as social interactions. Seemingly autonomous individuals and organizations in a social network are, in fact, embedded in social relations and interactions (Borgatti et al., 2009).

The perspective of social network analysis provides a set of methods for analyzing the structure of social entities as well as a variety of theories explaining the patterns observed in social networks (Wasserman and Galaskiewicz, 1994). Understanding the patterns that distinguish human behavior is of immense importance for deepening the knowledge about many ongoing phenomena such as spread of innovation or ideas, public health, group formation and information management, to name a few.

Social networks are fundamentally dynamic, and they evolve over time. Longitudinal social network data, i.e. time-event data is important, in order to assess whether the social embedding of an actor influenced the actor's behavior, or an actor's behavior resulted in change of relations. If social influence effects are present in the network then individuals are likely to change their attributes to conform to their friends (Raven, 1964). If social selection effects are present, then it is likely that individuals have a link to other individuals with similar attribute values. The consequence of these social phenomena is called homophily. Homophily means that a contact between similar people occurs at a higher rate than among dissimilar people. Thus, homophily potentially limits people's social space which has powerful implications on the information they receive, the attitudes they form, and the interactions they experience (McPherson, Smith-Lovin, and Cook, 2001). These social phenomena are shaped not just by the structure of social network but also depend on the position actors occupy within the network and how they interact.

In the last few years the interest in social network analysis has grown magnificently. This has primarily been triggered by the availability of data with exhaustive information of actor

interactions on a large scale. The world wide web, including mobile phones and online social networks have reshaped ways of communication and interaction, by providing the opportunity of being ubiquitously connected to everyone at any time. By their nature, these types of social interactions leave extensive digital traces of users' habits. For instance communication through mobile phones, online forums, emails and instant messaging documents our social interactions, location services provided by various social media applications capture our physical locations whereas, credit-card companies as well as E-commerce companies collect records of our online buying habits.

Since the last decade, online social networks (OSNs) have become an indispensable means of communication around the world. They have supplanted emails as the primary medium of sharing interesting information on the Internet (Benevenuto et al., 2009). They owe their success to taking cognizance of the predilection users have for ease with which they allow sharing information (pictures/videos/articles etc.) with their contacts; albeit, it is not clear how closely the interaction of users of an OSN resembles their interaction in the real world.

In this work we focus on the analysis of social ties and social interaction(with focus on OSNs). The understanding of the interplay between social relation, interaction, and attributes of actors could lead to a much better modeling of social networks. We analyze different topics related to interaction behavior and social networks analysis. The main goal of our work is to provide a better understanding of human interaction behavior when users are online, and further refine the modeling of social networks in order to improve the prediction of events and inference of links, and to determine group structure in online social networks. The work in this thesis combines analysis of large datasets from social media and communication networks, modeling and simulations, and predictive analysis on empirical data. We analyze a variety of OSNs data and engineer features that can help in getting a better understanding of the dynamics in these networks.

1.2. Organization

This dissertation is organized as follows:

Chapter 2 We introduce some definitions that are used in this thesis.

Chapter 3 The question of how to define the notion of a community has been an important focus of research. This chapter starts with covering various definitions of clusters/communities in a network. Clustering problems require partitioning a set of elements into homogeneous and well-separated subsets. Graph clustering is very hard which is intuitive at first sight but

is not very well defined. We give the background literature on community detection in social networks.

Chapter 4 This chapter covers the learning methods that are used in this thesis.

Rest of the thesis is divided into two parts.

Part I: In this part of the dissertation we analyze whether the social interaction patterns in OSNs reiterate and could refine the information about more persistent relations such as friendship ties. This part is divided into three chapters:

Chapter 5 We analyze interaction patterns when users are on an online social networking site. OSNs provide different types of personal and professional information sharing facilities which has led to their success as innovative social interaction platforms. In this chapter we analyze whether the persistent relations affect the instantaneous relations in online social networks. We study the interaction pattern of Facebook users and analyze whether the response of alters on each of the posts of ego depends upon the previous responses on the post or not, given the previous comments were from people belonging to the same or unknown community[1].

Chapter 6 We study the interplay between interaction and network structure. We first analyze the composition of overlapping social communities (circles) of an ego with respect to node attributes. Then, in a formative study we use the interaction information to obtain missing friendship ties[2].

Chapter 7 In this chapter we show the impact of additional interaction information on the inference of links between nodes in partially covert online social networks. In an elaborative study we show that interaction information can help infer unobserved (e.g. missing or hidden) social relations(friendship ties) more accurately. While privacy preserving mechanisms such as hiding one's friends list may be available to withhold personal information on online social networking sites, it is not overt whether or to which degree a user's social behavior renders such an attempt futile. Studies on link prediction have focused on properties such as existing network structure, actor attributes and interaction patterns to deduce information about the users. A major limitation of topology based features is observed when the network information is significantly missing which may lead to erroneous training set and eventually affect the

[1] *The research presented in this chapter is an extension of the work published in Nasim, Ilyas, et al. (2013)*

[2] *Parts of this chapter have previously been published in Nasim and Brandes (2014).*

performance of the classifier. In order to predict links in networks that are only partially observable, we utilized the stylized fact that individuals act as members of multiple social groups where members of the same group tend to participate in similar activities. Our results suggest that in the absence of network structure, interaction information may be used as a proxy to friendship ties and thus improves the performance of link prediction[3].

Part II: Sociological research has identified various dimensions of social relations, e.g., time, affect, intimacy, or reciprocal services (M. Granovetter, 1973) and group formation. In this part of the thesis, we study call logs data as an example of a pair-wise interaction.

Chapter 8 We analyze interaction from call logs data[4]. We explore how many active contacts do mobile phone users have; how often they are called; with respect to historic logs, we are interested in finding: Distribution of calls, more specifically, what percentage of communication goes to top contacts, and how often people call the recently called contacts.

Chapter 9 In the sociological context, most social interactions have fairly reliable temporal regularity. In this chapter we quantify the extension of this behavior to interactions on mobile phones. We expect that caller-callee interaction is not merely a result of randomness, rather it exhibits a temporal pattern. We first test the hypothesis that the majority of caller-callee interactions display temporal regularity. The model of user behavior assumed by call logs is, highly simplistic. It supposes that the likelihood of calling a particular contact, P(c), is a monotonically decreasing function of the time elapsed since last contact. Sociologists have, however, shown that human life is temporally organized and that most social interactions have fairly reliable temporal regularity. This implies that P(c) could be periodic. Such an implication, if correct, would allow for the design of a considerably more efficient calling interface than what is provided by either contact lists, or chronological call logs. To this end, we propose a call prediction model which takes into account the temporal calling patterns of users[5].

[3] *This chapter contains work from Nasim, Charbey, et al. (2016).*

[4] *Findings in this chapter will also appear in Nasim, Rextin, Khan, et al. (2016).*

[5] *Findings in this chapter are from Nasim, Rextin, Hayat, et al. (2017).*

2. Preliminaries

We begin with a set of essential definitions that will be used in this thesis.

Sociological Concepts [1]

Actors Actors are the basic unit of observation. In a socio-empirical study actors can be individuals (such as humans) or they can be aggregates (such as organizations).

Dyads and ties A pair of actors form a dyad, whereas, ties are data on dyads. A tie, is the union of all present or non-zero relationships of any particular ordered pair i and j.

Relation A relationship is a variable that is associated with a dyad. There are three aspects of such a variable: a content, a direction, and a value. A relation can thus be thought of as the entirety of all pairwise relationships that represent the same type of content.

Attribute An attribute is a collection of variables, each per actor.

Graph theoretic concepts

Graph A graph $G = (V, E)$ consists of a set of V vertices and a set of E edges that join pairs of vertices. Vertices also referred to as *nodes*. The vertex set and edge set of a graph G are denoted by $V(G)$ and $E(G)$ respectively. The cardinality of V is usually denoted by n and the cardinality of E is denoted by m. If two vertices are joined by an edge, then they are called *neighbors*. $(u, v) \in E$ is also referred to as e being incident on u and v or that u is adjacent to v.

A graph is called *undirected* if the vertex pair $\{u, v\} \in E$ is an unordered subset and *directed* if a vertex pair $(u, v) \in E$ is ordered. For a directed graph $G = (V, E)$, the underlying undirected graph is the undirected graph with vertex set V that has an undirected edge between two vertices $u, v \in V$ if (u, v) or (v, u) is in E.

The *neighborhood* $N(v)$ of a vertex $v \in V$ is the set of vertices that are adjacent to v.

[1] Hennig et al., 2012

Adjacency Matrix For a graph $G = (V,E)$, the adjacency matrix(x_{ij}), where $1 \leq i,j \leq |V|$ is defined by:

$$x_{ij} = \begin{cases} 1, & \text{if } (i,j) \in E \\ 0, & \text{otherwise} \end{cases}$$

Multigraphs If the edge set E contains the same edge several times, then E is a multiset. If an edge occurs several times in E, the copies of that edge are called parallel edges. Graphs that have parallel edges are also called multigraphs. A graph is simple, if each of its edges in contained in E at most once, i.e., if the graph does not have parallel edges. An edge joining a vertex to itself, is called a loop. In general, we assume all graphs to be loopless unless specified otherwise.

Induced subgraph A graph $H = (V',E')$ is a subgraph of the graph $G = (V,E)$ if $V' \subseteq V$ and $E' \subseteq E$. In vertex induced subgraph, E' contains all edges $e \in E$ that join vertices in V'. Thus the induced subgraph of $G = V,E$ with vertex set $V' \subseteq V$ is denoted by $G[V']$.

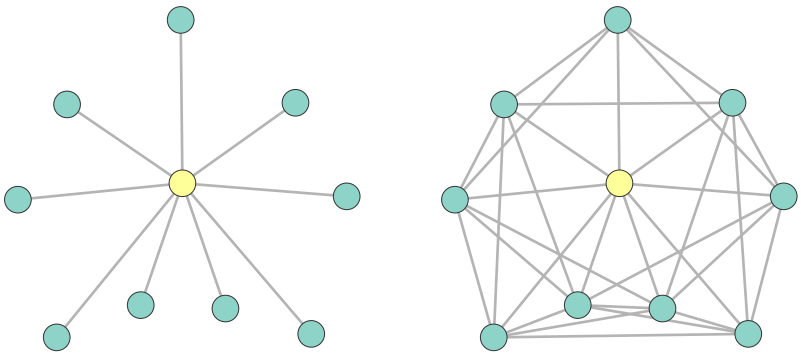
In edge induced subgraph, the edge set $E' \subseteq E$ is denoted by $G[E']$ is the subgraph $H = (V',E')$ of G, where V' is the set of all vertices in V that are joined by at least on edge in E'.

An edge will connect two vertices in the induced subgraph if and only if it was present in the original graph.

Walk, path and cycle A *walk* from a vertex x_0 to x_k in a graph $G = (V,E)$ is a sequence, $x_0,e_1,x_1,e_2,x_2,...,x_{k-1},e_k,x_k$, alternating between edges of G. The walk is called a *path* if $x_i \neq x_j$ for $i \neq j$. The length of a path is the number of vertices in the path. A walk with $x_0 = x_k$ is called a *cycle* if $e_i \neq e_j$ for $i \neq j$. In this thesis we denote the chordless cycles and paths on k vertices as C_k and P_k respectively. P_3 is a path on three vertices, whereas, C_3 is a cycle on three vertices.

Clique and isolates *Clique* is a subset of vertices of an undirected graph such that its induced subgraph is complete which means that all vertices in the clique are adjacent. An *isolated* vertex is a vertex with degree zero.

Ego and personal networks Ego and personal networks can be differentiated based on how the actors are embedded in social relations. Networks that describe a direct relation of an ego with the alters are the ego-centered networks (*ego-alter* dyads). Personal networks, in addition to the direct relation between ego and alters, also cover the relations between the alters (*ego-alter* and *alter-alter* dyads), Figure 2.1.



(a)An ego network. (b)A personal network.

Figure 2.1.: Ego and personal networks, an example.

Two-mode network and one mode projection A one mode network includes the relationships between actors of the same type, whereas a two-mode network includes the relationships that exist between two sets of units(for instance people or events).

3. Community Detection

Since Euler's solution to the Königsberg's bridges puzzle (Euler, 1741), a lot has been learnt about the mathematical properties of graphs (Bollobás, 2013). Graphs have been used for the representation of biological, technological, communication as well as social networks. In contrast to random graphs, real world networks such as structural representations of social networks, display inhomogeneities in the context of distribution of neighbors of a vertex, which is also known as the degree of the node. These inhomogeneities are not only limited to the global structure of the network, but, are also observed locally with high concentration of edges within certain vertices (or groups) and low concentration of edges outside the groups. This property of real networks is called *community structure* or *cluster*.

Given a graph G, a community can be thought of as a cohesive subgraph C, whose vertices are densely connected. The question of how to define the notion of a community has been an important focus of research. Cohesion of vertices in a graph can be quantified in several ways. The most strict *local* definition is based on the idea of a *clique*, which requires a complete subgraph. A clique is a maximal complete subgraph of two or more nodes such that nodes in the clique are all adjacent to each other but are not adjacent to any other node in the graph. These graphs are also known as cluster graphs. They form a hereditary class of graphs which can be characterized as P_3-free graphs. Definition of a community as a clique is a very stringent condition. Nonetheless, it is possible to relax this definition. Various generalizations of this definition exist in the literature. One possibility is to use properties related to the existence/non-existence of paths or cycles between vertices. For instance an *n-clique* is a maximal subgraph where the distance between each pair of vertices is not larger than n (Luce, 1950), (Alba, 1973). Mokken (1979) proposed two other alternatives, the *n-clan*, which is an n-clique whose diameter is not greater than n; and *n-club*, which is a maximal subgraph of diameter n. One of the generalizations of cluster graph through local structure is known as quasi-threshold graphs. In the case of cluster-editing, communities are found by finding a closest P_3-free graph, whereas, in the case of quasi-threshold graph one looks for a closest (P_4, C_4)-free graph.

Adjacency of vertices has also been mentioned as a criterion for subgraph cohesion which means that a vertex must be adjacent to some minimum number of other vertices in the subgraph. For instance, a *k-plex* is a maximal subgraph where each vertex is adjacent to all vertices of

the subgraph except at most k of them (Seidman and Foster, 1978). Another way to express cohesion of vertices in social network analysis is through *k-core*, which is a maximal subgraph where each vertex is adjacent to at least k other vertices in the subgraph (Seidman, 1983). These definitions foist conditions on both the minimal number of *absent* or *present* edges.

A cohesive subgraph can hardly be called a community if there is a strong cohesion not only between the vertices in the subgraph but also between the rest of the graph. It is imperative to compare the internal vs. external cohesion of the subgraph. An example of such a definition stems from social network analysis called *LS-set* or *strong community* (Radicchi et al., 2004). The idea is that the internal degree of each vertex in the subgraph is greater than its external degree.

Many methods are found in the literature which were developed to identify dense clusters/communities in networks. Graph partitioning methods have abundantly been used for community detection. Methods such as 'Minimum-cut' removes multiple edges at once that results in a hierarchical decomposition of components of a network (Zachary, 1977). Other graph partitioning methods include *Kernighan-Lin algorithm* (Kernighan and S. Lin, 1970), *spectral bisection method* (Barnes, 1982), *level structural partitioning*, *geometric algorithms*, etc. A description of these methods can be found in Pothen (1997).

The most popular class of methods to detect communities in graphs is based on the modularity based approach (Fortunato, 2010). The assumption behind these methods is that high values of modularity is indicative of good partitions, which may not be true in general. An example of modularity based method is clustering a graph using Girvan-Newman method (Girvan and Newman, 2002). Their algorithm is an example of a method that uses edge deletions for partitioning the network. Several other methods similar to Girvan-Newman have been suggested in literature. For instance Radicchi et al. (2004) observed that by removing edges that appear in few triangles (K_3), the result is similar to what is found by Girvan-Newman method.

Most of the methods that use the structural definition of community result in problem formulations that are $\mathcal{NP}$-complete (Nastos, 2015). There are various ways that can be used to extract structures in a network for instance using fixed-parameter tractability algorithm-technique (FPT). Several approximation algorithms have been designed for clustering networks that work by modifying edges i.e., they aim at minimizing the inter-cluster edges and maximizing intra-cluster edges.

The goal behind edge modification problems is to alter the edge set of a given graph as little as possible, in order to convert the given graph into a new graph that satisfies certain properties. Edge modification problems have a lot of application in many areas and recently have been studied in the context of detecting communities in social networks. Edge modification problems include completion, deletion and editing problems.

Table 3.1.: Complexity results for some edge modification problems. (Burzyn, Bonomo, and Durán, 2006), (Nastos and Gao, 2013), (Yunlong Liu et al., 2012), (Drange et al., 2015)

Graph Class	Completion	Deletion	Editing
Perfect	$\mathcal{NPC}$	$\mathcal{NPC}$	$\mathcal{NPC}$
Chordal	$\mathcal{NPC}$	$\mathcal{NPC}$	$\mathcal{NPC}$
Interval	$\mathcal{NPC}$	$\mathcal{NPC}$	$\mathcal{NPC}$
Chain	$\mathcal{NPC}$	$\mathcal{NPC}$	unknown
Comparability	$\mathcal{NPC}$	$\mathcal{NPC}$	$\mathcal{NPC}$
Cograph	$\mathcal{NPC}$	$\mathcal{NPC}$	$\mathcal{NPC}$
Threshold	$\mathcal{NPC}$	$\mathcal{NPC}$	$\mathcal{NPC}$
Bipartite	irrelevant	$\mathcal{NPC}$	$\mathcal{NPC}$
Split	$\mathcal{NPC}$	$\mathcal{NPC}$	$\mathcal{P}$
Cluster	$\mathcal{P}$	$\mathcal{NPC}$	$\mathcal{NPC}$
Quasi Threshold	$\mathcal{NPC}$	$\mathcal{NPC}$	$\mathcal{NPC}$

Let $G = (V,E)$ be a given graph. Consider a graph property Π, for instance the property defines the graph to belong to a certain graph class. For a given integer k, the Π-editing problem is to find the existence of a set of F unordered pairs of vertices such that $|F| \leq k$ and the resulting graph $G' = (V, E \triangle F)$ satisfies Π. The Π-deletion problem allows only edge deletions i.e., $F \subseteq E$ and the Π-completion problem allows only the addition of edges i.e., $F \cap E = \emptyset$.

There are various applications of edge modification problems. Edge modification has been studied in the context of physical mapping in molecular biology and human genome mapping (Bodlaender and Fluiter, 1996), (P. W. Goldberg et al., 1995).

The computational complexity of edge modification has been widely studied in the literature. Edge modification problems also constitute a broad range of $\mathcal{NP}$-complete problems. A summary of complexity results of some edge modification problems are provided in Table 3.1.

Overlapping Communities

Fortunato (2010), and Xie, Kelley, and Szymanski (2013) have reviewed a wide range of overlapping community detection algorithms along with reviewing several quality measures for the communities and existing benchmarks. We briefly cover some of the work done in detecting overlapping communities in networks.

Clique Percolation Method (CPM): CPM (Derényi, Palla, and Vicsek, 2005) is based on the

assumption that community consists of completely connected subgraphs which are overlapping. The algorithm begins by identifying all cliques of size k. A new graph is then constructed where each clique is represented by a vertex. Two vertices are connected if the k-cliques that represent them share $k-1$ vertices. Clique percolation method is suitable for finding overlapping communities in dense graphs. An example of overlapping communities is shown in Figure 3.1.

Link Clustering(LC): In Link Clustering links are partitioned instead of nodes. In (Ahn, Bagrow, and Lehmann, 2010), links are partitioned using hierarchical clustering of edge similarity. Given a pair of edges e_{ij} and e_{ik} that are incident on vertex i, the similarity can be computed using Jaccard coefficient as follows:

$$S(e_{ij}, e_{ik}) = \frac{|N(j) \cap N(k)|}{|N(j) \cup N(k)|} \tag{3.1}$$

A dendrogram is built using single-linkage hierarchical clustering. This dendrogram is cut at a threshold yielding link communities.

Mixed Membership Stochastic Blockmodels(MMSB): Airoldi et al. (2009) proposed mixed membership stochastic blockmodels which are a class of variance allocation models for pairwise measurements. MMSB provide exploratory tools for analyses in applications where the data can be represented as a collection of one-mode graphs. The nested variational inference algorithm is parallelizable. It allows fast approximate inference on large graphs.

Community-Affiliation Graph Model(AGM): J. Yang and Leskovec (2012) proposed Affiliation Graph Model (AGM). The graph model can generate synthetic networks and can also detect overlapping communities. The graph model is very similar to the model of Lattanzi and Sivakumar (2009) which suggested that the edge creation probability decreases with community size. However, AGM relaxes this assumption and allows arbitrarily large probabilities for edge creation, irrespective of the community size. This assumption is based on the previous work by Leskovec, Jon Kleinberg, and Faloutsos (2005) (the ratio of edges to vertices increases over time) and Leskovec, Backstrom, et al. (2008) (edges are created based on the principle of preferential attachment and by randomly closing triangles). These properties run counter to wisdom and also inconsistent with the previously proposed graph models. Authors showed the superiority of AGM over CPM, LC and MMSB both on synthetic data as well as on real-world data with known community structure. Additional proposed benefits of using AGM is the automatic estimation of number of communities in the network, unlike CPM or MMSB, which require the number of communities as input parameter.

Connected Iterative Scan: M. Goldberg et al. (2010) proposed an algorithm for analyzing the community structure of a large blog network using the interaction information between users. They used an undirected network representing user comments on blogs. From this bipartite

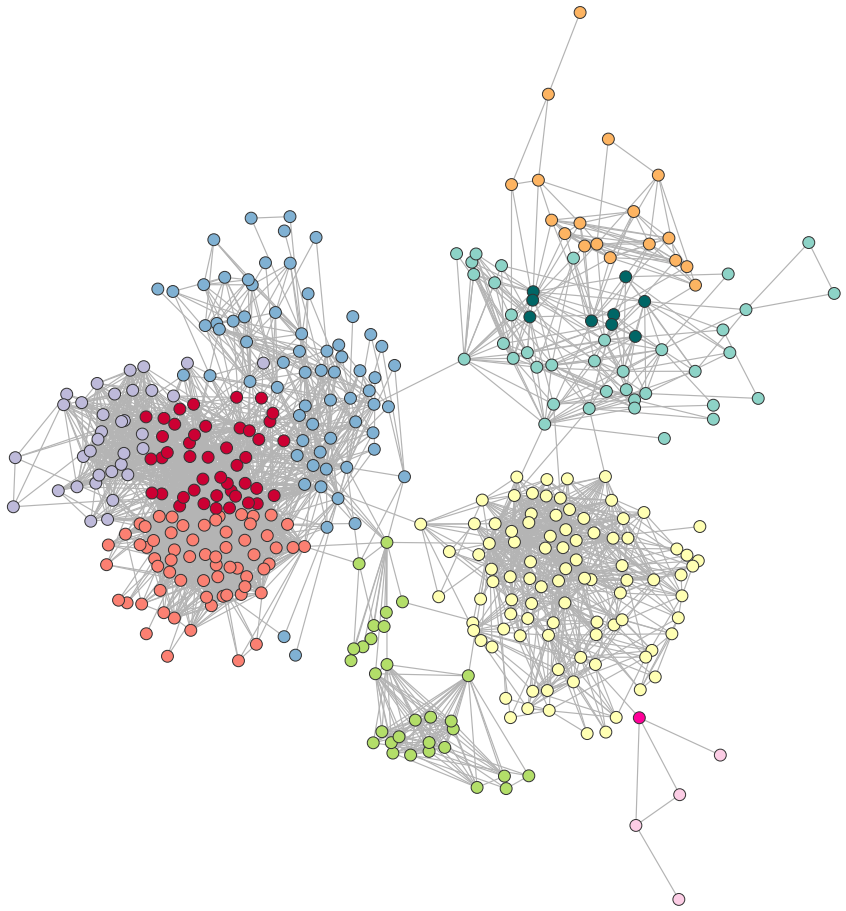


Figure 3.1.: This graph is an example of overlapping communities in the personal network of an ego. The dark colored nodes belong to more than one neighboring community. The communities were computed using clique percolation method.

network they create a friendship network. For instance, the number of times *A* writes a comment in response to a post by *B*, determines the weight of the edge shared by A and B which makes it a directed weighted network. They also checked whether group validity and overlap validity are satisfied for a given community or pair of communities.

4. Learning Methods

In this chapter we are going to introduce the methods which are used in this thesis for data analysis.

4.1. Classification

Machine learning methods are widely used in many applications and the most significant of those applications is *data mining* (Domingos, 2012). Programs can automatically be learned through machine learning systems. Applications of machine learning exist in recommender systems, anomaly detection, spam filters and web search to name a few. In machine learning, a *classifier* or a *learner* is a system that typically inputs a vector of continuous, categorical or binary *features* and outputs a single discrete value known as the *class*. An example of a classifier is a spam filter that classifies email into *spam* or *not spam* and its input can be a binary vector $\mathbf{x} = (x_1, ..., x_j, ..., x_d)$ where $x_j = 1$ if the j^{th} word in the dictionary is present in the email otherwise $x_j = 0$ (Domingos, 2012).

A set of examples/observations $(\mathbf{x_i}, y_i)$ called the *training set* is given as an input to the *learner*. Here $\mathbf{x_i} = (x_{i1}, ..., x_{id})$ is an observed input and y_i is the corresponding output, also known as *class label*. The learner consists of three main components: representation, evaluation and optimization. Learner can be divided into two types: ones where representation has a fixed size such as logistic regression and the ones where representation can grow with data such as decision trees.

Machine learning can broadly be classified into supervised learning and unsupervised learning. In supervised learning, observations are given with known label as compared to unsupervised learning where the observations are not labeled.

We will now have a look at four supervised learning algorithms that are used for analyzing data in this thesis. A review on classification techniques for supervised machine learning can be found in Kotsiantis, Zaharakis, and Pintelas (2007)

Naive Bayes

The most well representative statistical algorithms are the Bayesian networks. A comprehensive book on Bayesian networks is Jensen (1996). Naive Bayesian Networks (NB) are simple Bayesian networks that constitute directed acyclic graphs with the unobserved represented by only one parent and observed nodes are represented by the child nodes, with a strong assumption of independence among children (Good, 1950). Decision trees classify instances by sorting them based on feature values. A node in a decision tree represents a feature. In turn each branch represents a value that the node can assume. Instances are classified starting at the root node. They are sorted based on their feature values. Naive Bayes model is based on the estimating the following (Nilsson, 1965):

$$R = \frac{P(i|X)}{P(j|X)} = \frac{P(i)\prod P(X_r|i)}{P(j)\prod P(X_r|j)} \tag{4.1}$$

The larger of the two probabilities indicate the class label that is likely to be the actual label (if $R > 1$, i is predicted else j is predicted).

Naive Bayes classifiers have an underlying assumption of independence among child nodes. This assumption is not always true, therefore naive Bayes classifier is less accurate than other sophisticated algorithms. The major advantage of naive Bayes classifier lies in its short computation time for training set. Additionally, the model has the form of a product which can be converted to sum through logarithms which can give significant computational benefits.

Logistic Regression

The difference between logistic regression and linear regression models is that the class label in logistic regression is binary or dichotomous.

In regression problems the important quantity is the mean value of the outcome variable that is also known as *conditional mean*, $E(Y|x)$, where Y is the outcome binary variable and x is the value of the explanatory variable. $E(Y|x)$ is read as the expected value of Y given x. When Y is a dichotomous variable then $E[Y|x]$ represents the conditional probability that Y value 1 given the value of x, i.e., $E[Y|x] = P[Y = 1|x]$. Shortly we will denote this probability as $\pi(x)$. In linear regression the assumption is that the mean can be expressed as a linear equation (Hosmer Jr and Lemeshow, 2004):

$$E(Y|x) = \beta_0 + \beta_1 x \tag{4.2}$$

This implies that $E(Y|x)$ can possibly take any value since the range for x is between $-\infty$ and ∞. For dichotomous data, the conditional mean must be greater than or equal to zero and

less than or equal to one. Therefore, a linear model as in equation 4.2is not adequate to model binary data, and a link function $g(x)$ that transforms the interval $[0,1]$ into the real line $(-\infty,\infty)$ must be used.

A number of distribution functions have been considered for analyzing dichotomous outcome variable. The two main reasons to choose logistic distribution lies in the fact that is extremely flexible, an easily used function and allows meaningful interpretation of data.

Let $\pi(x) = E(Y|x)$ be the conditional mean of Y given x when logistic distribution is used. Let the following equation represent the logistic regression model:

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} \tag{4.3}$$

Several link functions have been proposed, among them the logit function, defined as:

$$\begin{aligned} g(x) &= ln\left[\frac{\pi(x)}{1-\pi(x)}\right] \\ &= \beta_0+\beta_1 x \end{aligned} \tag{4.4}$$

This formula can be easily generalized for multivariate case. The logit of the multiple regression model where there are p independent variables, $\mathbf{x}' = x_1, x_2, \ldots, x_p$, is given by:

$$g(\mathbf{x}) = \beta_0+\beta_1 x_1+\cdots+\beta_p x_p \tag{4.5}$$

in this case the logistic regression model is given by:

$$\pi(\mathbf{x}) = \frac{e^{g(x)}}{1+e^{g(x)}} \tag{4.6}$$

Logit transformation is important because $g(x)$ has many of the desired properties of a linear regression model. Further, $g(x)$ is linear, may be continuous. It ranges from $-\infty$ to ∞ depending upon the range of x.

Lets assume we have n independent observations for the pair (x_i, y_i), where $i = 1,2,\ldots,n$ and y_i denotes the value of the binary variable for the i^{th} subject. Further, assume that the binary outcome is coded as 1 or 0, representing the presence or absence of a characteristic.

In order to fit a logistic regression model to a set of data, it is required to estimate the parameters, β_0 and β_1. These parameters are estimated using *Maximum likelihood Estimation* (MLE).

MLE is a common learning algorithm used by a variety of machine learning algorithms for estimating the parameters of a statistical model.

The interpretation of regression coefficients β is along the same lines as in linear models. The left hand side of the equation is a logit rather than a mean. Change in the logit of the probability

associated with a unit change in the j_{th} predictor holding all other predictors constant, is represented by β_j.

If Y is coded as 0 or 1 then $\pi(x)$ Equation 4.3 provides the conditional probability for $Y = 1$ given x, denoted as $P(Y = 1|x)$, and $1 - \pi(x)$ gives the conditional probability that $Y = 0$, given x, denoted by $P(Y = 0|x)$. An easy way to express the contribution of the pair (x_i, y_i), to the likelihood function is as follows:

$$\pi(x_i)^{y_i}[1 = \pi(x_i)]^{1-y} \tag{4.7}$$

The likelihood function is the product of the terms given in Equation 4.7 because the observations are assumed to be independent.

$$l(\beta) = \prod_{i}^{n} \pi(x_i)^{y_i}[1 - \pi(x_i)]^{1-y} \tag{4.8}$$

According to MLE, we use a value of β that maximizes the expression in Equation 4.8. This expression can be expressed as a *log likelihood* function:

$$L(\beta) = ln[l(\beta)] = \sum_{i=1}^{n} \{y_i \, ln[\pi(x_i)] + (1-y)ln[1 - \pi(x_i)]\} \tag{4.9}$$

Differentiating $L(\beta)$ w.r.t. β_0 and β_1 and set the resulting expressions equal to zero we find the value of β that maximizes $L(\beta)$.

$$\sum_{i=1}^{m} [y - \pi(x_i)] = 0 \tag{4.10}$$

$$\sum_{i=1}^{m} x_i[y - \pi(x_i)] = 0 \tag{4.11}$$

where m is the number of observations.

The expressions in Equations 4.10 and 4.11 are non linear in β_0 and β_1 and require special methods for their solution.

The value of β given by Equations 4.10 and 4.11 is called the maximum likelihood estimate denoted by $\hat{\beta}$. It provides an estimate of the $P(Y = 1|x = x_i)$. It represents the predicted value for the logistic regression model. A consequence of Equation 4.10 is that the sum of the observed values of y is equal to the sum of the expected values of y:

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{\pi}(x_i) \tag{4.12}$$

Linear Discriminant Analysis

Despite Logistic regression being a simple and powerful linear classification algorithm, it has its limitations. One of the limitations of logistic regression is the *two -class* problem. The algorithm is intended for binary classification problems. It can be extended for multi-class classification, but is not often used for this purpose. Logistic regression may become unstable when the classes are well separated, as well as in the case when there are few examples for estimating the parameters.

Linear Discriminant Analysis (LDA) does addresses the limitations of logistic regression. It is useful for multiclass classification and even for binary-classification problems, it is a good idea to try both logistic regression and LDAs.

Linear discriminant analysis, assumes that cases of a each class k are generated according to some probabilities (π_k) and the predictor variables are generated by a class-specific multivariate normal distribution.

Given a number of independent features LDA creates a linear combination of the features that yield the largest mean differences between the desired classes.

For simplicity lets assume there are two classes in the dataset. The mean of each class (μ_1 and μ_2) and mean of entire dataset (μ_3) is computed (Balakrishnama and Ganapathiraju, 1998):

$$\mu_3 = p_1\mu_1 + p_2\mu_2 \tag{4.13}$$

where p_1 and p_2 are the apriori probabilities of classes and in the simplest case assumed to be 0.5.

The class separability is determined based upon the within-class and between-class scatter, which is computed as follows:

$$S_w = \sum_j p_j(cov_j) \tag{4.14}$$

The covariance matrices are symmetric and computed using the following equation

$$cov_j = (x_j - \mu_j)(x_j - \mu_j)^T \tag{4.15}$$

The between-class scatter is computed as follows:

$$S_b = \sum_j (\mu_j - \mu_3)(\mu_j - \mu_3)^T \tag{4.16}$$

The optimization criterion in LDA is the ratio of the between-class scatter to the within-class scatter. The axes of the transformed space are defined by the maximizing this criterion.

The Eigen vector of a transformation in a $1-D$ invariant subspace of the vector space in which the transformation is being applied. Any vector space can be represented in terms of linear combination of eigen vectors. For a K class problem there are $K-1$ non-zero eigen values.

For the class depended LDA,

$$transformed - set_j = transform_j^T Xset_j \tag{4.17}$$

For the class independent LDA

$$transformed - set = transform - spec^T Xdataset^T \tag{4.18}$$

The test vectors are transformed and classified using eh Euclidean distance. Once LDA transformation are completed, Euclidean or Root Mean Square distance is used to classify data points. For n classes, n Euclidean distances are obtained for each observation. The smallest Euclidean distance classifies the observation's predicted class.

LDA can be described as prototype method, where each class is represented by a prototype; cases are assigned the class with the nearest prototype.

Logistic regression is an alternative to Fisher's 1936 method, linear discriminant analysis (LDA), however, logistic regression does not require the multivariate normal assumption of LDA.

Support Vector Machines

Support Vector Machines (SVMs) revolve around the concept of a *margin* - either side of a hyperplane that separates two classes. The idea is to maximize the margin, hence creating the maximum possible distance between the separating hyperplane (see Figure 4.1). The instances that lie on either side of the hyperplane have proven to reduce an upper bound of the expected generalization error.

In the case of a linearly separable training data, a pair $(\mathbf{w}, b)$ exists such that (Kotsiantis, Zaharakis, and Pintelas, 2007),

$$\mathbf{w^T x_i} + b \geq 1, \text{ for all } \mathbf{x_i} \in P$$

$$\mathbf{w^T x_i} + b \leq -1, \text{ for all } \mathbf{x_i} \in N$$

The decision rule is determined by $f_{\mathbf{w},b}(\mathbf{x}) = sgn(\mathbf{w^T x} + b)$. Here $\mathbf{w}$ is the weight vector and b is the bias.

If the data is linearly separable, an optimum separating hyperplane can be determined by minimizing the squared norm of the separating hyperplane. This step can be described as a convex quadratic programing problem:

$$\underset{\mathbf{w},b}{\text{Minimize}}\ \Phi(\mathbf{w}) = \frac{1}{2}||\mathbf{w}||^2 \tag{4.19}$$

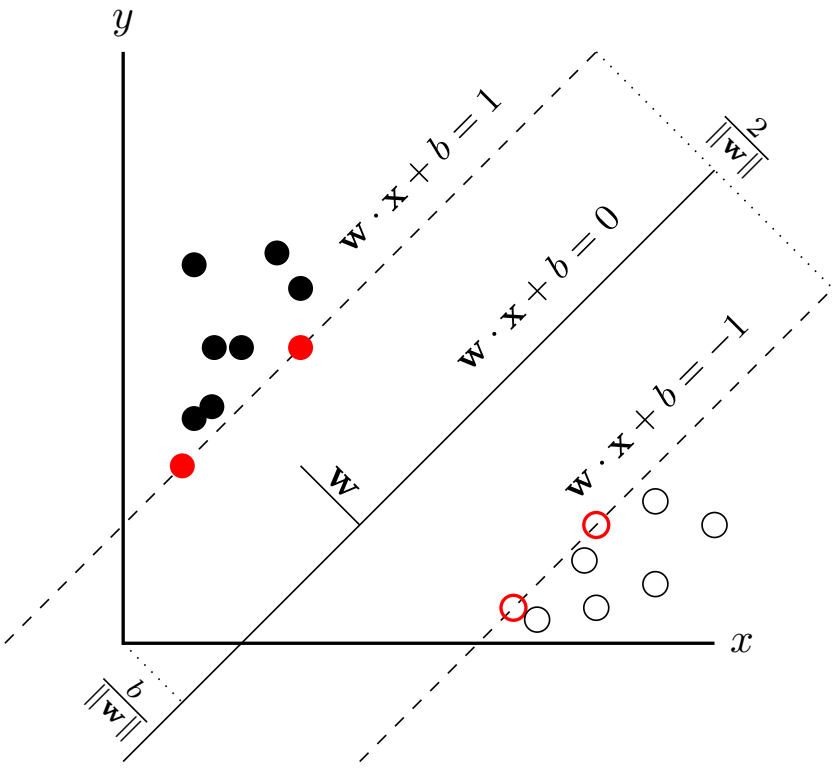


Figure 4.1.: Maximum margin in SVMs for two classes.

Data points lying on the margin of the optimum separating hyperplane are known as support vector points. The linear combination of support vector points form solution set and the other points are ignored. For this reason, SVMs are well suited to the tasks where number of features are large since the number of support vectors selected by the model is usually small.

When the data contains misclassified instances, the classifier may not be able to find any separating hyperplane. *Soft margin* can help mitigate this problem by accepting some misclassifications of training instances (Veropoulos, Campbell, and Cristianini, 1999). This is achieved by introducing positive slack variables ξ_i, where $i = 1,...,N$ in the constraints. Therefore,

$$w.x_i - b \geq +1 - \xi \text{ for } y_i = +1$$

$$w.x_i - b \leq -1 + \xi \text{ for } y_i = -1$$

$$\xi \geq 0$$

An error can occur if and only if the corresponding ξ_i exceed unity. The training error is bounded above by $\sum_{\mathbf{i}} \xi_i$. In this case the Lagrangian is:

$$L_p \equiv \frac{1}{2}||w||^2 + C\sum_i \xi_i - \sum_i \alpha_i\{y_i(x_i.w - b) - 1 + \xi_i\} - \sum_i \mu_i \xi_i \tag{4.20}$$

Here μ_i are the Lagrange multipliers in order to enforce positivity of the slack variable ξ_i.

For real world problems it is common that they involve non-separable data which means that no hyperplane exists that may successfully separates the instances in the training set. A common way to mitigate this problem is through mapping the data into a higher-dimensional space which is called *transformed feature space*.

When chosen appropriately, a transformed feature space can make any consistent training set separable. A linear separation in the transformation space reflects a non-linear separation in the original *input space*. When the data is mapped to some Hilbert space H, as $\Phi : R^d \rightarrow H$, with possibly infinite dimensions, the classifier would only depend on the data through dot products in H of the form $\Phi(x_i).\Phi(x_j)$.

Had there been a kernel function K, such that $K(x_i,x_j) = \Phi(x_i).\Phi(x_j)$, then one would use K in the training algorithm. Thus, Φ would never be determined. Kernels allow inner products to be calculated directly in feature space without performing the mapping as discussed earlier. After the creation of a hyperplane, new points are mapped to the feature space through kernel functions. This entails a careful selection of a kernel function since it defines the transformed feature space where training instance are classified (Genton, 2001). A common practice is to estimate on a range of settings and then by doing cross-validation, find the best one. This contributes to the slow speed of SVMs.

Some popular kernels are:

1. $K(x,y) = (x.y+1)^P$
2. $K(x,y) = e^{\frac{-||x-y||^2}{2\sigma^2}}$
3. $K(x,y) = tanh(\kappa x.y - \Sigma)^P$

The computations required by SVM are very time consuming since it requires solving N^{th} dimensional quadratic programming problem, where N is the number of samples in the training dataset. An algorithm called Sequential Minimal Optimization, decomposes the problem into smaller quadratic programming problems, thus it can solve the SVM QP relatively quickly.

In general support vector machines perform well when features are continuous. Logic based systems perform better for dealing with discrete/categorical features (Kotsiantis, Zaharakis, and Pintelas, 2007).

A survey on SVMs can be found in Burges (1998).

4.2. Model Considerations

When there is insufficient data to train the model, one can run into a problem called *overfitting*. For instance when the learner outputs a classifier that is 100% accurate on training data but only 50% accurate on test data, when it should have output a classifier that is 75% accurate on both training and test data (Domingos, 2012). Overfitting comes in many forms, and can be understood by generalizing the error into *bias* and *variance*(Domingos, 2000). Bias is a measure of the contribution to error of the central tendency of the classifier i.e., consistency to learn the same wrong thing. Variance measures the contribution of error of deviations from the central tendency, i.e., tendency to learn random things irrespective of the real signal. Algorithms with a high-bias tend to generate simple yet highly constrained models that are insensitive to data fluctuations, in order to keep the variance low. Naive Bayes assumes that the dataset is from a single probability distribution, thus, it is considered to have a high bias. Algorithms that have a high-variance profile can generate complex models that fit the data fluctuations more readily. Such models include SVMs, decision trees and neural networks. High-variance models are prone to overfitting.

Cross-validation technique can help against overfitting. It is a model validation technique to assess how the results will generalize to an independent data set. Cross-validation involves partitioning a sample of data into complementary subsets and learning the model on one subset call training and validating the analysis on the other subset called testing. In order to reduce variability, multiple rounds (such as 10-folds) of cross-validation are recommended. A popular method to combat overfitting is by adding a *regularization* term to the evaluation function. Another common option is to use a statistical significance test such as χ^2-test. One of the common misconception about overfitting is that it is a result of noise in the training examples. However, in a case where we learn a Boolean classifier which is only the disjunction of the examples labeled true in the training set; in this case the classifier gets the no noise in the training set but would output every positive test example wrong.

Apart overfitting, another common problem with machine learning is known as *curse of dimensionality*. Many algorithms work well in low dimensions but become intractable when the input is high-dimensional i.e., consists of a large number of features. Generalizing becomes harder as the dimensions increase. If there are irrelevant features, the noise from them hide the correct signal and the model may end up making random predictions.

The most important factor in machine learning projects is the features. When features are independent and they correlate well with the class, then learning is easy. Machine learning is an iterative process of learning, analyzing and modifying the data or learner. Creating *model ensembles* is a now a standard way of improving the results (Bauer and Kohavi, 1999). There are three main was to use model ensembles: *bagging*, *boosting* and *stacking*. In *bagging* one

simply generates variations of the training set and learn a classifier on each variation. Results are combined by voting. This method reduces the variance. In *boosting*, training examples have weights which can be varied so that the new classifier focuses on the observations that were learned wrong by the previous classifiers. Finally, *stacking* uses the outputs of individual classifiers as inputs for higher level learner.

4.3. Expectation Maximization Algorithm for Data Clustering

Finite mixture models can be used to model the distribution and clustering of a wide variety of data. In this thesis we consider their application in the context of clustering.

Let $\mathbf{y}$ be a *p-dimensional* vector denoted by $\mathbf{y} = (y_1, \ldots, y_p)^T$, contains the values of p variables measured on each of n entities that are to be clustered. Let $\mathbf{y}_j$ be the value of $\mathbf{y}$ corresponding to the jth entity, where $j = 1, \ldots, n$. Let $\mathbf{y}_1, \ldots, \mathbf{y_n}$ be the observed random sample from a mixture of finite number, say c, of groups in some unknown proportions $\pi_1, \ldots, \pi_c$.

Mixture density of $\mathbf{y}_j$ can be represented as follows (X. Wu et al., 2008):

$$f(y_i; \Psi) = \sum_{i=1}^{c} \pi_i f_i(y_j; \theta_i) \ (j = 1, \ldots, n) \tag{4.21}$$

Here, $\pi_1, \ldots, \pi_c$ sum to one and the conditional density $f_i(y_j; \theta_i)$ is specified up to a vector θ_i of the unknown parameters i.e., $i = 1, \ldots, c$. All the unknown parameters can be represented by a vector transpose as follows:

$$\Psi = (\pi_1, \ldots, \pi_{g-1}, \theta_1^T, \ldots, \theta_c^T)^T \tag{4.22}$$

A probabilistic clustering of data into c clusters can be achieved in terms of the posterior probabilities of component membership,

$$\tau_i(y_j, \Psi) = \frac{\pi_i f_i(y_j; \theta_i}{f(y_j; \Psi)} \tag{4.23}$$

where, $\tau_i(\mathbf{y}_j)$ is the posterior probability that y_j belongs to the ith component of the mixture.

Ψ can be estimated by MLE. The estimates of Ψ and $\hat{\Psi}$ is given by an appropriate root of the MLE equation,

$$\frac{\delta \log L(\Psi)}{\delta \Psi} = 0 \tag{4.24}$$

where the log likelihood function for Ψ is given by,

$$\log L(\Psi) = \sum_{j=1}^{n} \log f(y_i;\Psi) \tag{4.25}$$

Solutions of 4.25 corresponding to local maximizers can be obtained via the EM algorithm.

An advantage of adopting mixture models with elliptically symmetric components is that the clustering process is independent of the irrelevant factors such as orientation of clusters in space.

The E and M steps of the EM algorithm for the MLE estimation of multivariate normal components are described in (G. McLachlan and Peel, 2004). In the EM framework the unobservable component labels are z_{ij}, which is the missing data and z_{ij} is defined to be one or zero depending whether y_{ij} has membership in the ith component of the mixture $(i = 1,\ldots,c; j = 1,\ldots,n)$.

At the $(k+1)$th iteration, the estimation step requires taking the expectation of the complete data log likelihood $\log L_c(\Psi)$, given the current estimate Ψ^k for Ψ. The E-step is affected by replace z_{ij} by $\tau_{ij}^{(}k)$ which is the posterior probability that y_j belongs to the ith component of the mixture.

The current fit is expressed as:

$$\tau_{ij}^{(}k) = \frac{\pi_i^{(}k)\phi(y_j;\mu_i^{(k)},\sum_i^k)}{f(y_j;\Psi^{(k)})} \tag{4.26}$$

The maximization the updated estimates π_j, the mean vector μ_j and the covariance matrix $\sum_i$ for the ith component are given by:

$$\pi_i^{(k+1)} = \frac{\sum_{j=1}^{n} \tau_{ij}^{(k)}}{n} \tag{4.27}$$

$$\mu_i^{(k+1)} = \frac{\sum_{j=1}^{n} \tau_{ij}^{(k)} y_j}{\sum_{j=1}^{n} \tau_{ij}^{(k)}} \tag{4.28}$$

and

$$\sum_{i}^{(k+1)} = \frac{\sum_{j=1}^{n} \tau_{ij}^{(k)} (y_i - \mu_i^{(k+1)})(y_j - \mu_i^{(k+1)})^T}{\sum_{j=1}^{n} \tau_{ij}^{(k)}} \tag{4.29}$$

The expectation and maximization alternate till the change in estimated parameters of log likelihood are less than a certain threshold.

The choice of value of c, i.e., the number of clusters can be made by performing likelihood ratio test or by using Bayesian Information criterion.

Part I.

Relations and Interaction

5. Interplay Between Social Communities and Interaction

Relations between people account for many aspects of communication, such as credibility, empathy, attraction, etc., and ultimately the effectiveness of communication (Rogers and Bhowmik, 1970). In the *Relational analysis* approach the unit of analysis is pairs of individuals and the relationship between them.

> *"The distinctive characteristic of networks is that their units of observation (the identifiers of data points) are not single entities but pairs of entities, and that each entity may appear in multiple such pairs"*, (Brandes, Robins, et al., 2013).

The most fundamental principle of human communication is the exchange of messages. According to the sociological theories on social interaction, this exchange of messages most frequently occurs between an initiator and a receiver who are homophilous, i.e., alike or similar. Homophily refers to the extent to which interacting pairs of individuals are similar with respect to some attributes, such as beliefs, values, education, social status, etc. Homophily limits people's social space. This has implications on the information they receive, the attitudes they form, and the interactions they experience (McPherson, Smith-Lovin, and Cook, 2001). On the other hand, heterophily is the degree to which interacting pairs of individuals are different with respect to certain attributes.

According to Tarde (1903):

> *"Social relations, I repeat, are much closer between individuals who resemble each other in occupation and education."*

In a free-choice situation, when a person can interact with any one of the various possible people, there is a strong disposition for him/her to select a person who is like himself/herself. Many sociological studies provide empirical evidence of the homophily principle in communications context. For instance the political influence patterns in a presidential election were homophilous with respect to age and social status (Lazarsfeld, Berelson, and Gaudet, 1968). Interactions among members of a legislature were between those of equal age, partisanship,

and prestige (Wahlke, 1962). In another study it was observed that Chicago ghetto dwellers shared family planning ideas with others of similar social status, age, marital status, and family size (Palmore, 1967).

One possible reason attributed to frequent communication between homophilous individuals is the effectiveness in communication when the source(s) and receiver(s) share common meanings, attitudes, and beliefs. Thus, communication between them is likely to be more effective. More often than not, individuals cherish the comfort of communicating with others who are similar in social status, education, beliefs, etc. On the other hand, heterophilic interaction is likely to cause message distortion, and may cause cognitive dissonance, an uncomfortable psychological state, as the receiver is exposed to messages that may be inconsistent with his existing beliefs and attitudes (Rogers and Bhowmik, 1970). Another reason for an increased effectiveness in communication of homophilous sources is that such homophily leads to greater credibility. In communication, credibility is the degree to which a source is perceived as *trustworthy*. In a study on a traditional Indian village, it was found that peasants associated a high credibility to their fellow villagers. However, with the advent of new communication channels in the village, the system was transformed to a more open system, and the qualification credibility shifted to agricultural scientists, extension agents, and radio. Nevertheless, the safety credibility remained with homophilous peers(Rogers and Bhowmik, 1970).

How behavior is affected by the social relations is described as one of the classic questions of social theory (Mark Granovetter, 1985). Garfinkel (1948/2005) studied the social world with respect to *actor*, *situation*, *group* and *time*. He illustrated how *"conversation aligns the interactants' inner sense of time or 'inner duree', creating a 'new time dimension' and a 'common vivid presence'"*. A listener who is not an active part of the conversation, experiences the occurrences of the others action as *events occurring in outer time and space*. Notwithstandingly, at the same time he also experiences his observatory actions as a sequence of retentions and anticipations happening in his own inner time in order to understand 'message' as a meaningful unit. The communicator's speech, as well as the listener's vivid presence both occur simultaneously. Hence a new time dimension comes into being. Both the parties can later claim that they experienced the occurrence of an event together. Garfinkel's account depicts that instead of physical co-presence, merely shared time, is relevant to identify an event as a shared event. He showed no interest in locations as physical places rather he emphasized that places are situations and their coherence is given by the practices which constitute them. This view on shared events is a motivating factor for studying interactions which do not share physical presence such as interaction in the online space.

In the next section we move on to look at the dynamics of interaction in online social

networks (OSNs)[1].

5.1. Online Social Networks

The Internet spans several types of information sharing systems, including the world wide web. In the last decade, online social networks have gained significant popularity. They have become an indispensable means of communication around the world. For instance, Facebook (over 1.59 billion users), LinkedIn (over 414 million), LiveJournal (over 10 million), and Pinterest(over 100 million users) are popular social networking sites. According to a research report by Nielsen Online, Americans spent more than quarter of their online time on social networking sites (*Nielsen* 2012). They spent 121 billion minutes on online social networks between July 2011 and July 2012; Facebook was the most-visited social networking site during that time. In the fourth quarter of 2013, on average, smartphone owners spent nearly 11 hours per person, accessing social networking and search applications. These statistics show that off late online social networks have started having a critical bearing on the lives of Internet users. Online social networks have supplanted emails as the primary medium of sharing interesting information on the Internet (Benevenuto et al., 2009). They owe their success to taking cognizance of the predilection users have for ease with which they allow sharing information (pictures/videos/articles etc.) with their contacts. Content on online social networks is organized around users. Participating users join a social networking site; they can then publish their profile, and can create links to other users which can be friends, acquaintances, professional contacts or complete strangers. The consequential social network provides a way to maintain social connections, or find users with similar interests (Mislove, Marcon, et al., 2007).

An in-depth understanding of the dynamics in online social networks which includes the underlying network structure, attributes of users as well as interaction patterns, is necessary to evaluate current systems, and to design future online social networks based systems. Online social networks are usually run by individual corporations (e.g. Google, Facebook and Yahoo!), and are accessible via the Web and Smartphone applications. They are already at the heart of some very popular Web sites. It is likely that online social networks will play a critical role in future online interaction, be it at commercial or personal level. These networks offer an unparalleled opportunity for Sociologist to study social networks at a large scale. They can examine the data to test existing theories about social networks, and can look for new forms of behavior found in online social networks. Studying online social networks can also help in improving online viral marketing. Discovering idiosyncrasies in the online space is

[1]Please note that online social networks and OSNs are used interchangeably in this thesis.

critical for comprehending the security of these networks. Understanding nuances and overlaps in user behavior when they are on a social networking site is important because it can lead to better site design, help understand social interaction and can be helpful in designing the next-generation Internet infrastructure and content distribution systems (Benevenuto et al., 2009). Notwithstanding one's position on these phenomena, a better understanding of the structure of, and dynamics behind online social networks odds-on to improve our understanding of the benefits, shortcomings, and perils associated with these ideas (Mislove, Marcon, et al., 2007).

5.1.1. Interaction on Facebook

Facebook (*Facebook* 2016) is an online social networking service based in Menlo Park, California, United States. The social networking website was launched on February 4, 2004 by Mark Zuckerberg at Harvard. Initially the website was limited to Harvard students, but later expanded for other higher education institutions and since 2006, anyone aged 13 and older is allowed to become a registered user of the website. Once users are registered, they can create a user profile and add other users as 'friends'. Users can exchange messages with friends, post status updates and photos, write comments, share videos, use various Facebook/third-party apps designed for this website, like pages, and receive notifications when friends for fan pages update their profiles.

In the next section, we reflect on the homophilous behavior of users on online social networking sites, by capturing the online social interactions on Facebook. The case study was published in (Nasim, Ilyas, et al., 2013).

5.2. Case Study: Commenting Behavior of Facebook Users

Contribution

The contribution of this case-study is two folds

1. Collection of an original Facebook dataset[2].
2. A usual means of interaction on Facebook is by posting status messages on one's own wall. Friends then comment on these status updates in a comments thread. We propose that Facebook users exhibit dependency in commenting behavior. Their commenting

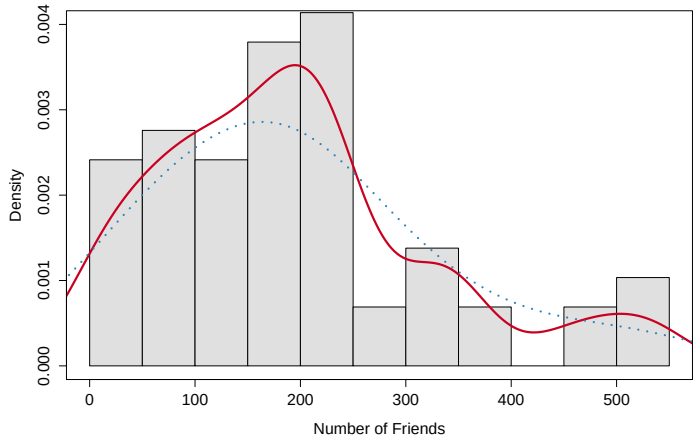
[2]I collected this dataset in January 2012 while working at NUST School of Electrical Engineering and Computer Science, Islamabad, Pakistan

behavior is more aligned with their friends or in other words it is autocorrelated with their homophilic group of friends.

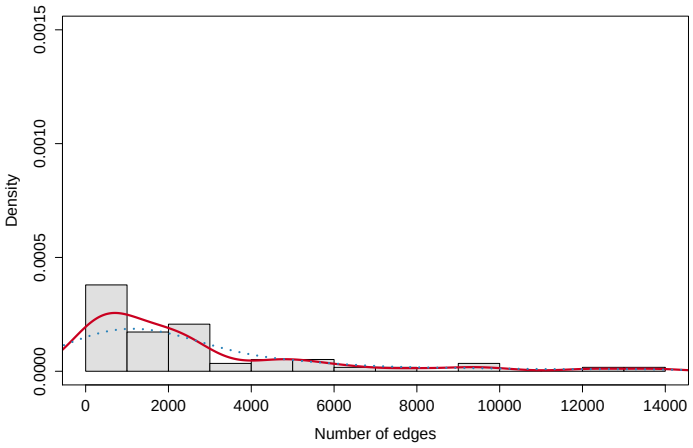
For analysis purpose, we divide friends in communities based on their mutual friendships (obtained from ego network of users) and study the sequence of comments on users' status posts. Ego networks are divided into communities because possibly homophily is the sociological force behind clustering of similar individuals, thus this segregation helps to capture commenting behavior of related friends.

We collected a dataset of wall posts of 50 volunteer Facebook users for this study. Facebook users add close friends, professional connections and sometimes strangers as well. Facebook does not distinguish between these different types of friends. However, in real life these friends can be divided into different social circles or groups on the basis of family, school and organizational ties. Facebook content sharing has received significant attention in recent past (Y. Liu et al., 2011), (Christofides, Muise, and Desmarais, 2009). Facebook provides different interaction options to its users. It allows users to manage access on every piece of content shared (*Facebook* 2016). The user requirements on OSNs entails this flexibility since content sharing decisions may vary from type of content.

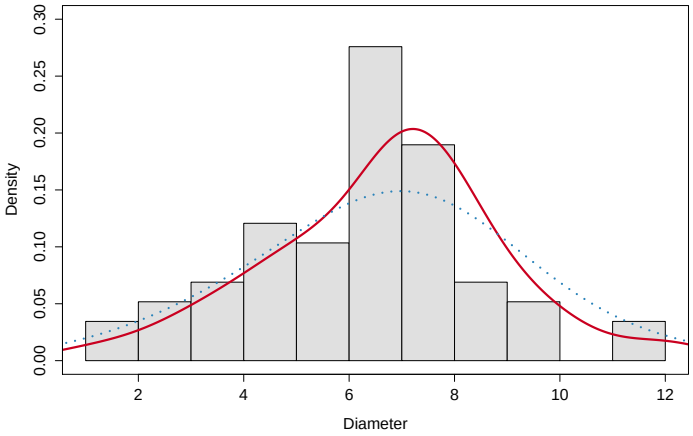
The next section describes our approach for collecting data from Facebook users to analyze their commenting behavior. We then provide some basic statistics about the dataset.



(a)Distribution of friends

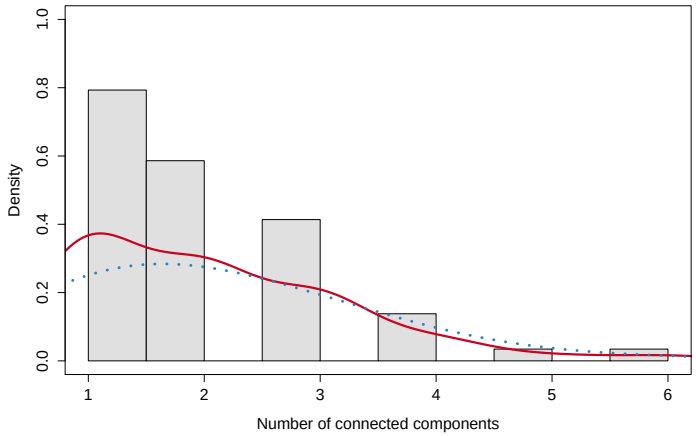


(b)Distribution of edges

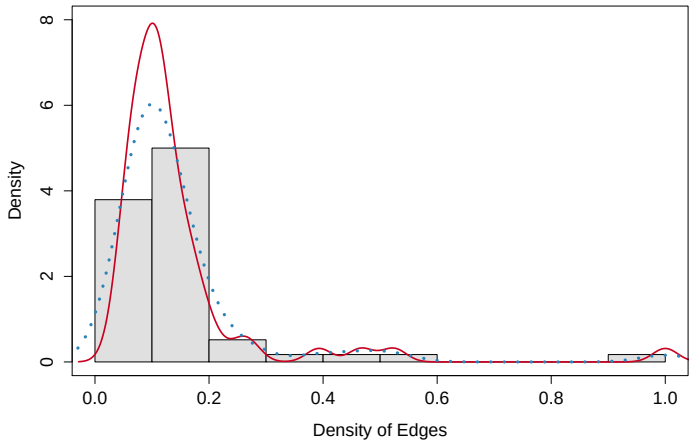


(c)Distribution of diamteter

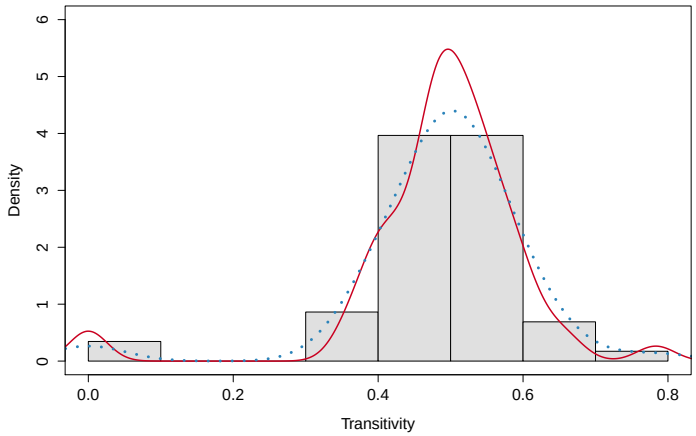
Figure 5.1.: Network Statistics



(a)Distribution of connected components



(b)Distribution of edge density



(c)Distribution of transitivity

Figure 5.2.: Network Statistics

5.2.1. Dataset

Data Collection We developed a Facebook survey application that was distributed to a group of volunteer Facebook users. Participants were asked to allow the survey application to collect information about their past status posts. Permissions include access to friends, status posts and related content. Our application is called "Facebook Content Sharing Survey." It retrieves the identities of friends and the order in which they commented on retrievable status messages [3]. At this point, a user has the option to quit the application if he / she has apprehensions about giving access to private data. Once the application is installed, the user is requested to fill a given survey. At the end of the survey, the application posts a message on the user's wall. This encouraged friends of the user to take the same survey. Once the application acquires the permissions it retrieves data which consists of friend IDs, status message IDs, comment IDs and information about friendship connections between users' friends. This data is hashed and anonymized before writing to disk.

Dataset Statistics The Facebook application was deployed on January 3, 2012; 58 users filled the survey between then and January 5, 2012.

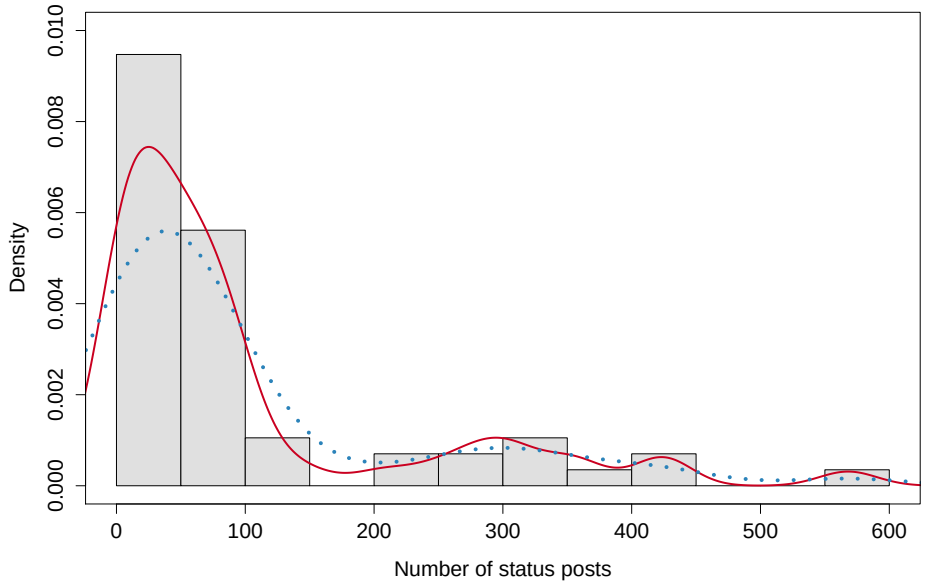
On average, users in this dataset had 116 status updates with an average of 7 comments per status and an average of 193 friends and 2500 edges in the network. Figure 5.1 shows the distribution of different network statistics.The average diameter of each network was 6.7 and the average number of connected components were 2.06.

Collectively, the dataset comprises of 5,778 status updates and 40,186 comments. Eight profiles had insufficient network/commenting information. Figure 5.3 shows the distribution of status posts and of total comments. Around 20% profiles have more than 100 status posts. About 80% of profiles have less than 1000 comments.

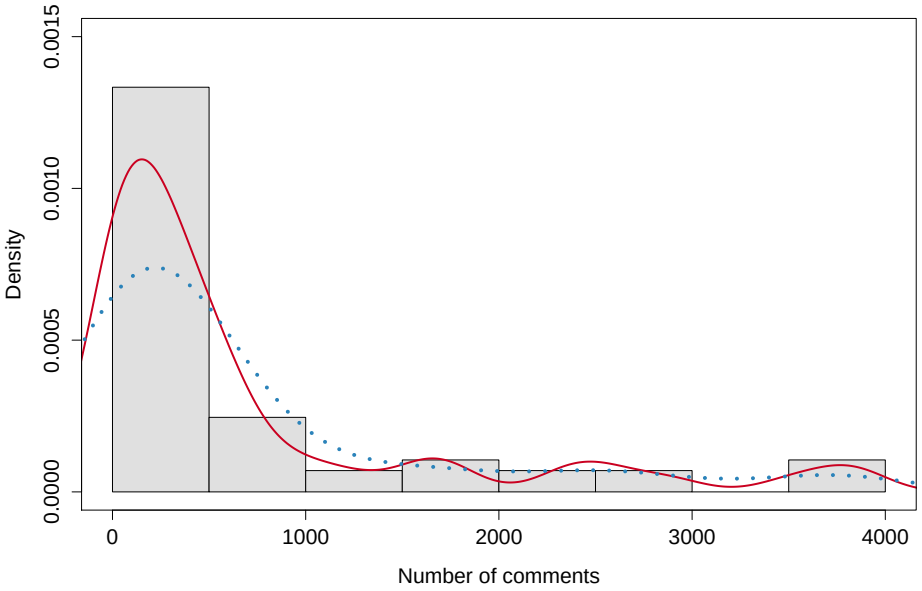
The acquired data helped us study the commenting behavior of approximately 8,500 Facebook users.

The Survey asked the participants to answer whether they would like to comment on a post of their friend if someone unknown to them has already commented on that post. It also asked the users whether they are encouraged to comment on a friend's status message when someone known to them has already commented and vice versa. Participants were given three choices: 1) Yes, 2) No and 3) Never thought about it. Users were also asked to write a reason for their chosen answer. Users were distributed over different geographical locations.

[3] The app does not read contents of status messages or comments.



(a)Distribution of status posts



(b)Distribution of total comments

Figure 5.3.: Status posts statistics.

5.2.2. Methodology

As discussed earlier, people are acquainted with a diverse group of friends. Facebook treats these diverse groups of people as 'friends'. It is not clear how closely the interaction of users of an OSN resembles their interaction in the real world. This question was motivated by the observation that we have different groups of friends (childhood school friends, college friends, workplace colleagues) and that the dynamics of real life interaction with friends of these different groups is often very different.

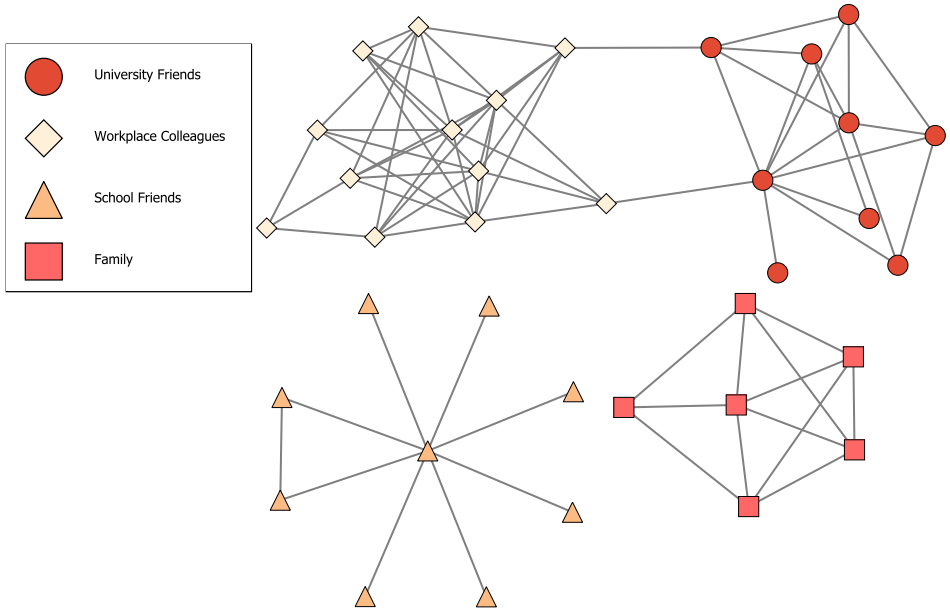
People change their behavior to align it with the behavior of their friends (Easley and J. Kleinberg, 2010). This process has been described as *social influence* (Raven, 1964),(Moscovici, Sherrard, and Heinz, 1976),(Friedkin, 2006) since the social connections in the friendship network of the user are influencing the individual characteristics of the user. Similarly, people with similar attributes tend to form friendship links which is described as *social selection* (Kandel, 1978). If social influence effects are present in the network then individuals are likely to change their attributes to conform to their friends. If social selection effects are present, then it is likely that individuals have a link to other individuals with similar attribute values. The consequence of these social phenomena is called homophily. Homophily means that a contact between similar people occurs at a higher rate than among dissimilar people. Thus, homophily potentially limits people's social space which has powerful implications on the information they receive, the attitudes they form, and the interactions they experience (McPherson, Smith-Lovin, and Cook, 2001).

In everyday life we notice that most people hangout with their family, college and school friends separately since people's personal networks are homogeneous in terms of many socio-demographic, behavioral, and intra-personal characteristics. In this section, we analyze the dataset we collected by means of our 'Facebook Content Sharing' application, to see if users' homophily in their social interactions extend to Facebook. We accomplish this by observing the commenting behavior of friends on a user's status post. We divide the friends of a user into communities and find the probability of each community commenting on a status message. Community detection tends to find groups of friends. For example, our old class mates are more likely to be friends with each other rather with our workplace colleagues.

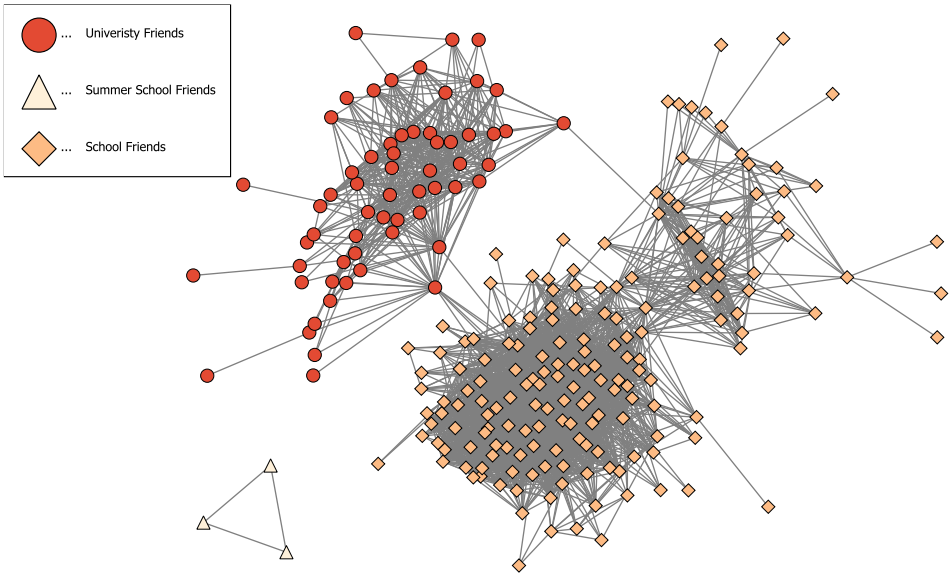
As a first step, for every participant, we build a graph of the relationships between the participant's friends (one-hop neighbors) called a mutual friendship graph (MFG).

An MFG is defined as follows:

Definition Let $G = (V, E)$ be a graph and $N(w) = \{x \in V | \exists \{x, w\} \in E\}$ the neighborhood of a node $w \in V$. Further, let $u, v \in V$.
Then the mutual friendship graph G' of u and v is induced by the set $V' = \{N(u) \cup N(v) \cup \{u, v\}\}$ on G.



(a)



(b)

Figure 5.4.: These are Mutual Friendship graphs of two users. The one shown in (a) has four communities with few inter-community links. The one in (b) has three communities with relatively more links across communities.

Since we are analyzing individual user profiles, we find communities in the personal network of each ego. We used Stanford Network Analysis Package (SNAP) (Leskovec and Krevl, 2014) which is a general purpose network analysis and graph mining library and *visone*(*visone* 2012), a tool for visual social network analysis, to cluster the personal networks. We find

communities using the Girvan-Newman (GN) (Girvan and Newman, 2002) and the Iterative Conductance Cutting (ICC) (Kannan, Vempala, and Vetta, 2004) algorithms. GN identifies communities based on the principle of betweenness. The ICC iteratively splits clusters using minimum conductance cuts. Finding a cut with minimum conductance is $NP-$hard. Therefore, a threshold input value α is used. Splitting of a cluster ends when the approximation value reaches the threshold. (For further insight on ICC, readers are referred to Brandes, Gaertler, and Wagner, 2003). For our dataset a higher value of α resulted in many small clusters with many inter-community links. We set the value of α to its suggested default value of 0.2 (provided in *visone*). In most cases, clustering through ICC results in fewer clusters compared to GN, with relatively few inter-community links. In Figure 5.4, examples of MFGs of two random users from the dataset are shown. Since the ego(user) is a friend of all of his/her friends, therefore not shown in the graph. These graph were derived using ICC and they show Facebook friends belonging to different social communities.

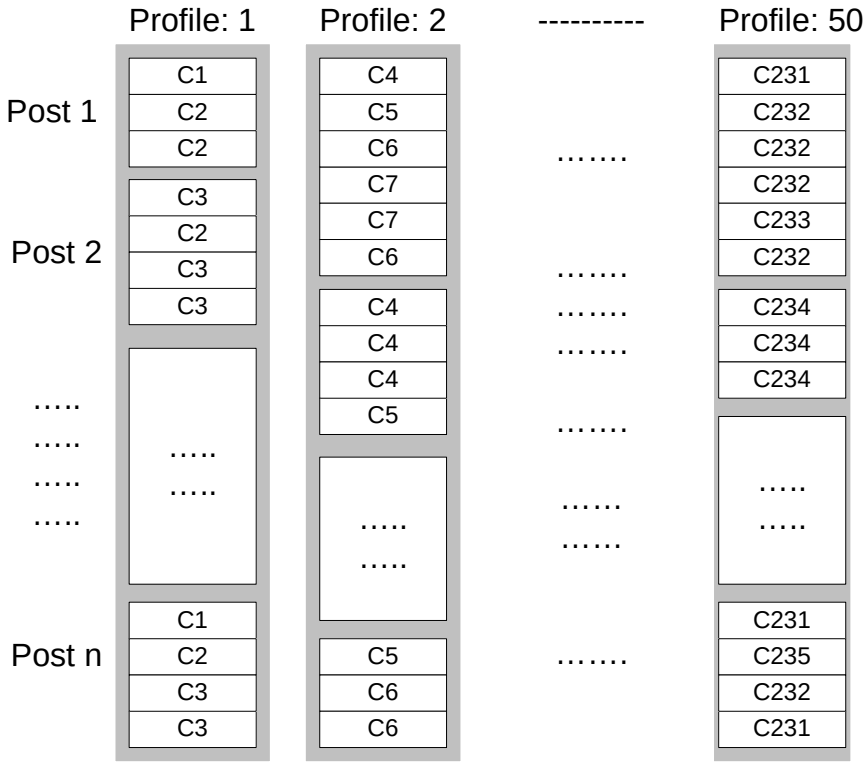


Figure 5.5.: There are 50 profiles shown as shaded rectangles. Each collected profile contains multiple status posts. Each status has comments from different communities. For example, the first post of Profile 1 has comments from communities $C1$, $C2$ and $C2$.

5.2.3. User Behavior Model

We model the decision of a Facebook friend belonging to a community of friends Ci by a Bernoulli random variable X, where $X = 0$ represents the outcome that a user does not comment on a post and $X = 1$ denotes the outcome that a user does comment.

For the understanding of the reader, we have provided a graphical illustration in Figure 5.5. Vertical columns shaded in grey show the profiles under analysis. Each profile contains multiple status posts and each status has comments from different communities. The comments on each post are grouped together. To determine whether or not a dependence exists in the online social interactions of Facebook users, we compute the probability density function (pdf) of X conditioned on two parameters:

A) The total number of comments a post has already received. This number is modeled by discrete random variable N.

B) The total number of comments a post has received from members of his / her own community. This number is modeled by discrete random variable Y.

This way, $p_X[X = 1|N = n, Y = y]$ denotes the probability that a user belonging to a particular community will comment ($X = 1$) on a status post that has until that point received n comments out of which y are from members of his / her own community, where $0 \leq y \leq n$. Obviously, the probability that a user does not comment ($X = 0$) under the same circumstances is $p_X[X = 0|N = n, Y = y]$. Using our collected dataset described above, we estimate these conditional pdfs for various valid combinations of Y and N. We also compute the 90% confidence interval for each of these conditional probabilities, modeled as a Gaussian pdf. The 90% confidence interval is represented by whiskers around the estimated probability in Figures 5.6 and 5.7.

5.2.4. Results

In this section we discuss results obtained from the Facebook dataset. We make three assumptions for our analysis:

1. Commenting behavior on Facebook to be independent of the specific community they are part of.
2. Facebook users' decision to comment or not comment on a post is independent of the nature of the post.
3. Each alter belongs to a single community, implying that there are no overlapping communities.

These assumptions are essential to the analysis since characteristics such as nature of the post cannot be captured without analyzing the content of the posts. We agree that some posts are

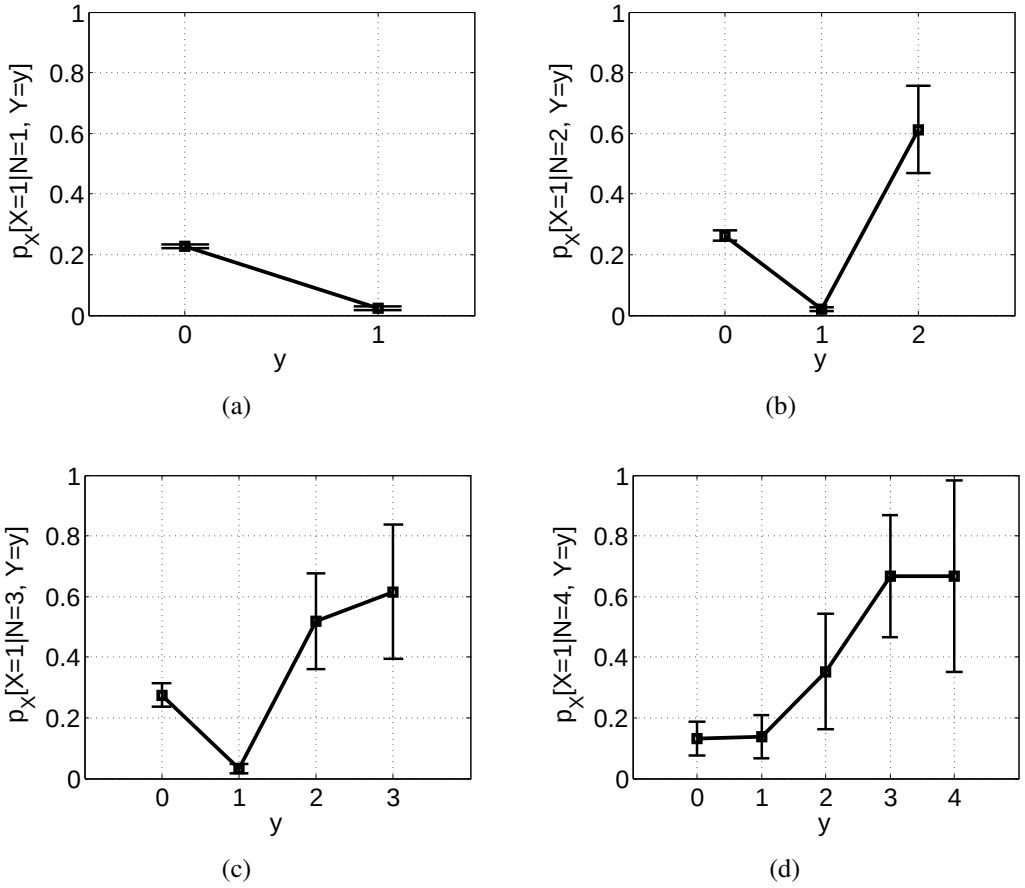


Figure 5.6.: Analysis of commenting behavior using Girvan-Newman Clustering for dividing the friends in communities. Number of comments (y) on a post is shown on x-axis. Probability of receiving the next comment from the same community is plotted as $p_X[X = 1|N = n, Y = y]$ on Y axis. Except for the case where total number of comments on a post is 1, the probability of receiving the next comment from the same community is increasing given the number of comments from the same community are increasing, as shown in (b) to (d).

more interesting than the others, however, for the sake of anonymization, the content of posts or comments was not recorded. It was also ensured that the analyzed profiles had status posts visible to all the friends.

Figures 5.6 and 5.7 depict the probabilities of a user commenting on a status post $p_X[X = 1|N = n, Y = y]$ using GN Clustering and ICC. Figures 5.6a and 5.7a plot $p_X[X = 1|N = 1, Y = 0, 1]$, the probability of commenting on a post that has already received one comment ($N = 1$). The graph shows the probability of receiving the next comment from a user when there are $N = 1$ comments on a status post out of which $Y = 0, 1$ are from his / her own community. In Figure 5.6a, based on the GN based community detection, the probability of commenting is higher when there is one comment from an out-of-community person than when the previous comment

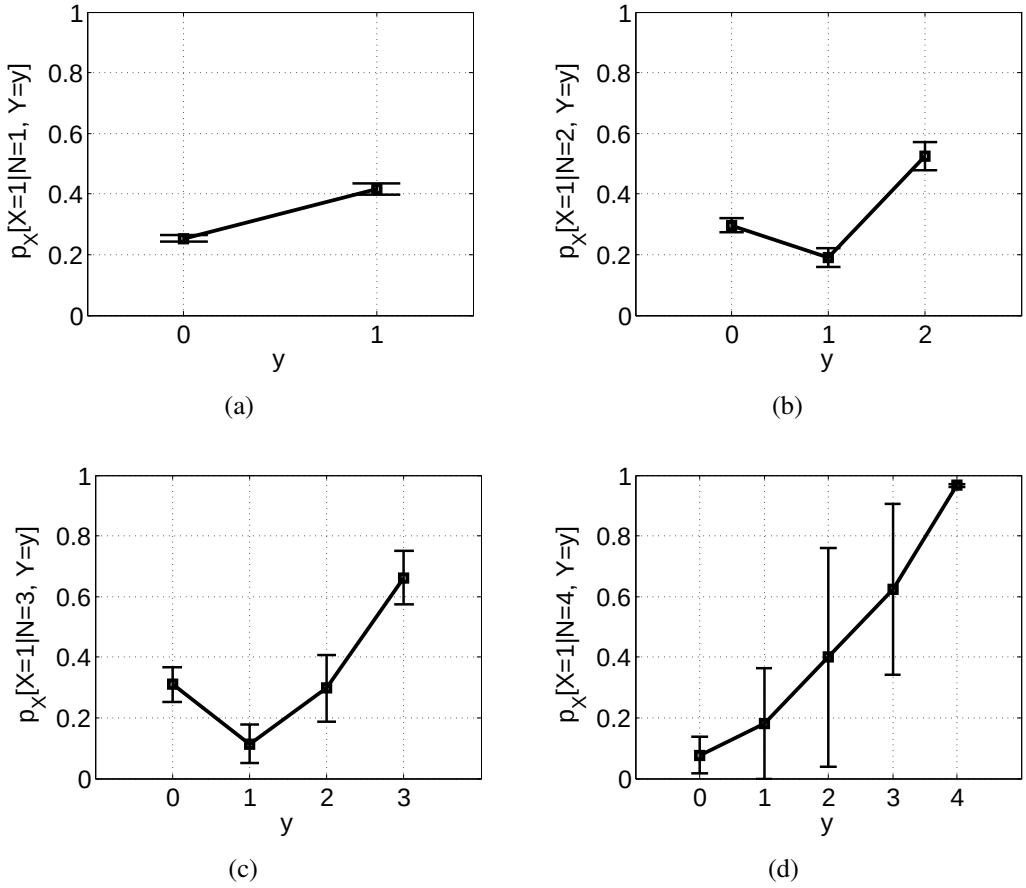


Figure 5.7.: Analysis of commenting behavior using Iterative Conductance Cutting for dividing the friends in communities. Number of comments (y) on a post is shown on x-axis. Probability of receiving the next comment from the same community is plotted as $p_X[X = 1|N = n, Y = y]$ on the vertical axis. The probability of receiving the next comment from the same community is increasing given the number of comments from the same community are increasing, as shown in (a) to (d).

is from within his / her own community, i.e. $p_X[X = 1|N = 1, Y = 0] > p_X[X = 1|N = 1, Y = 1]$. Interestingly, when community detection is performed using ICC, Figure 5.7a the opposite trend is observed, i.e. $p_X[X = 1|N = 1, Y = 0] < p_X[X = 1|N = 1, Y = 1]$. A single comment may have a subtle effect which can be a reason for $p_X[X = 1|N = 1, Y = 1]$ to be lower than $p_X[X = 1|N = 1, Y = 0]$ in Figure 5.6a. Since, performing clustering via ICC in Figure 5.7a the opposite trend is observed, therefore, this trend can also be attributed to the limitation of community detection scheme(GN) where many small clusters were found.

However, as N increases, $p_X[X = 1|N > 1, Y \leq n]$ an interesting trend begins to emerge. Figures 5.6b through 5.6d plot the probability of commenting when $N = 2, 3$ and 4 when the GN algorithm is used for community identification, i.e. $p_X[X = 1|N = 2, Y = 0, 1, 2]$, $p_X[X =$

$1|N=3,Y=0,1,2,3]$ and $p_X[X=1|N=4,Y=0,1,2,3,4]$. The same trend is observed in Figures 5.7b through 5.7d when ICC is used for $N=2,3,4$ with increasing values of Y. Since friends belonging to same community might share common traits and may tend to conform their behavior with their friends. A similar trend is reflected in commenting behavior. Even for the case where two friends of a user belong to same community but do not have a link between them, one in the presence of a comment from the other is encouraged to comment because they have many mutual friends. It is also important to mention that we are not considering the behavior of individuals, rather groups of people having ties between them.

The observations we made based on the quantitative results are also supported by the responses we received from survey participants. Some participants' responses for the following question, "Based on your experience please write your thoughts about this scenario: Your school friend posts a status update. You see comments on that status message from a number of people who are unknown to you. Would you be discouraged to write your comment on that post?" are:

> *"I wouldn't want to bother people I don't know about my comments or intrude on another group's conversation".*

> *"There are already a lot of people posting and my comment would probably get lost amongst them. I also don't want to see notifications when unknown people continue to comment on the post".*

Thus, the quantitative results appear to show a steady trend in Facebook users to comment on a status post with greater likelihood if more of the preceding comments are from people within their circle of friends.

5.3. Discussion

In this chapter we quantitatively analyzed the commenting behavior of Facebook users. We observed that friends belonging to a particular community of friends are more likely to comment on a status post if a large fraction of prior comments are by other Facebook users belonging to their own community, i.e. Facebook users tend to exhibit a dependence (on previous comments) in their decision to comment on a post based on previous comments. This means that the information propagating through Facebook tends to coalesce (to some extent) within communities. Our findings suggest that design improvements in visibility/non-visibility of comments to other communities may allow more user engagement and encourage user interaction on online social networking sites. The findings also suggest that social influence effects are present when users

are interacting online and that the individuals are likely to change their behavior to conform to their friends.

The idea of selective self-presentation is not new. It can be traced back to Goffman (1959), who described interactions between actors and their audience as,

"a performance in which some traits are accentuated while others are concealed".

Depending upon the situation, individuals make a series of decisions, concerning how to present themselves before people they are interacting with. Schlenker (1985) suggested three key factors behind self-presentation:context, audience, and environment. In another study Leary (1995) argued that individuals want either to conform to the values of the audience or they want to invoke a desired response. In these situations self-presentations may enhance their image. The unique features of computer mediated communication attenuate the process of self-presentation because individuals are able to carefully construct their self-presentation through textual asynchronous communication (Walther, 1996). Nevertheless, users still engage in selective self-presentation in a variety of ways when they are online (Zhao, Grasmuck, and Martin, 2008).

With the advent of OSNs, there is an overwhelming interest in content sharing behavior of the users. Most of the studies aim at identifying the various privacy needs in terms of content sharing. A study conducted by Ozenc and Farnham, 2011, on sharing of online content, reveals that people favor focused sharing. The study concluded that people organize their OSN space based on the nature of their relationships. Results highlighted the fact that people have too many connections called 'friends' and are inundated by online content. Users need a mechanism for improving relevancy in OSN.

Authors argue that focused sharing allows a greater degree of control over sharing which is not found in current privacy settings. The existing policy configuration tools are difficult for users to understand. Moreover, they do not always provide the desired content sharing settings.

Fang and LeFevre (2010) presented a template for the design of a privacy wizard to mitigate the burden of privacy management for users. The paper suggests that it is less burdensome for the users if their personal privacy preferences are automatically derived through some mechanism. The model generated from this template takes very little input from the user but configures the user's detailed privacy settings. To illustrate this idea authors built a sample wizard, based on an active learning paradigm. The experimental evaluation based on privacy preference information collected from 45 Facebook users, indicated that the community structure of a user's OSN is a valuable resource when modeling users' privacy preferences. Kairam et al. (2012) conducted research on selective sharing on Google+. Their analysis suggested that users engaging in selective sharing consider three main factors when choosing audiences for their content. It includes privacy, relevance and social norms. Selective sharing helps them

manage their self-presentation to different audiences. Users prefer to tailor their profiles to show different information to different social circles. This is also an indication that since at a time users are a part of different groups, they want to show that their interests are aligned with each group, thus they prefer selective sharing.

The study conducted on Facebook content sharing settings by Y. Liu et al. (2011), measured the disparity between the desired and actual content settings which quantifies the enormity of the problem of managing privacy. Authors conducted a survey on 200 Facebook users and found the overall privacy settings match the users' expectations only 37% of the time. In order to mitigate this problem, the authors support the suggestions of Fang and LeFevre, 2010 and propose that user created friends lists can be helpful in implementing new tools for managing content settings since there is significant correlation between the circle of friends and content settings.

Vitak (2012) examined multiple dimensions of users' public disclosures. They studied, status updates, establishment of relationship between users' network composition,privacy concerns, and characteristics of public updates. They concluded that specific qualities of one's disclosures on Facebook have a positive relation with their perceptions of social capital. In another study it was found that having too many online friends and access to different social capital derange the sharing process because the users have faced social surveillance and social control in the past. This kind of social control often forces younger people to use 'conformity as a strategy' when sharing content (pictures, information, comments) in order to maintain their privacy (Brandtzæg, Lüders, and Skjetne, 2010).

Although customized content sharing is gaining importance in OSNs, identification of various content sharing needs of users still needs further investigation. User behavior varies across different groups of friends. More specifically, there is a need to analyze that when certain content is exposed to two distinct groups of friends on an OSN, how does each group behave in the presence of the other.

One of the limitations we identified in our analysis was the choice of community detection algorithm. Since the community finding algorithms are heuristic based therefore, it cannot be said with certainty that they identify the optimal communities in all cases. For analysis purposes we also assumed that an alter belongs to at most one community. Although in reality it is likely that alters may belong to more than one community (social circle). Hence, in order to avoid limiting the analysis with one approach, we used two different algorithms. Further analysis can be performed with other community finding approaches. To our knowledge, this is the first study on the analysis of influence of prior comments on a status post in Facebook. Identification of subtle behavioral patterns such as friendship influence on commenting behavior is a precursor to further investigating the idiosyncrasies of focused content sharing of social media users. This

can help in improving the design of next generation of user interface for OSNs with the aim of enhancing online user experience.

In the this chapter we observed that network structure was a good predictor of future commenters. In the next chapters we are going to see whether interaction is a proxy to network structure.

6. Interaction and Social Relations

In this chapter we analyze three aspects of social networks, namely: relational ties, actor attributes and interaction among actors.

> *"A social network consists of a finite set or sets of actors and the relation or relations defined on them", (Wasserman and Galaskiewicz, 1994)*

.

The relationships defined on a set of dyads are the pair-wise relations defined on a set of actors. The dyads constitute a larger structure. The overlapping structure of dyads in the network may result in interdependence among the relationships (Hennig et al., 2012). Network analysis approaches current empirical social research in a way that not only analyzes the characteristics of individuals such as age, gender, education or status but it also analyzes how individuals act in different contexts and the roles in which they are embedded in their social environment (Hennig et al., 2012).

This chapter is organized as follows:

In Section 6.2, we analyze the composition (with respect to attributes of actors) of overlapping social communities, known as *social circles* in the sociological parlance.

In Section 6.3 we cluster the discussions from an online social networking site and map them on the core network. Further, in Section 6.4, we report the findings from a pretest that shows an empirical evidence on inferring social relations (in this case friendship ties), using interaction(commenting) information from Facebook.

6.1. Friendships and Foci

In the previous chapter we assumed that an alter belongs to at most one social group, whereas, in reality it is possibly not the case. We also noticed that our results were not invariant to the community detection algorithm, rather, we see a variation in results based on the underlying clustering algorithm. The work in this chapter is motivated by the fact that individuals may belong to multiple communities or social circles. According to Collins dictionary, a social

circle is, *"a group of people who are socially connected"*. The interaction of individuals in the society and their friendship ties are helpful in assigning them social circles (Simmel, 1908).

According to sociological theories of friendships and group formation, the reason behind friendships lies in the shared interests, personal preferences, or ascribed status of the individuals who participate in joint activities (Breiger, 1974). Similarity brings people together because individuals organize relations around points of common interest known as 'Foci' (Feld, 1981). A focus is defined as a social, psychological, legal or physical entity around which joint activities are organized, e.g., workplaces, voluntary organizations, hangouts, families, etc. When a focus is more constrained, there is more interaction expected. The amount of time 'pairs' of individuals spend in activities associated with a particular focus, indicates the constraint of that focus (Feld, 1981). The structural approach underlying the focus theory suggests that when there are restrictions on time, effort or emotions then individuals will experience certain pressure to combine their interactions with different members of their network. This is done by finding and developing new foci around which similar people are brought together. This will be expedited if the foci upon which the original ties are based are more "compatible" that is, involve similar types of activities and social interactions (e.g., childhood friends and family are typically more compatible with each other than workplace and childhood friends). When the focus theory is extended to OSNs, it is observed that on OSNs, people who are friends with each other share similar interests. For instance they join similar groups (e.g. on Flickr, Facebook etc.), like similar pages (e.g., on Facebook) and follow same people (e.g., on Facebook, Google+, Twitter etc.) or involve in same discussions.

The challenging task is to ascertain which individuals are associated with which foci and investigate the constraint size and compatibility of important foci since individuals form relations around many different foci (Kadushin, 1966). In *The Web of Group Affiliations* (Simmel, 1903), Simmel suggests that an individual is a part of sufficient circles which have *form* and *organization*. Membership in these circles gives an individual the opportunity to pursue each of his interests in association with others. The specific qualities of the individual are preserved through the combination of circles.

Much of the focus on community detection in social networks has been laid on identifying disjoint communities. Lack of a unified definition of a community makes it difficult to compare the performance of community detection methods. Much of the work has focused on hierarchical partitioning problem. This approach is prudent for some types of networks for instance taxonomies or organizational networks, but in social networks individuals are seen as sharing focus around family, workplace, organizations etc. In turn many social networks contain overlapping communities since members are attached to multiple foci.

In online social networking sites, people can connect to several social groups. In recent

years researchers have focused on finding the circles of users who are active on online social networking sites (Leroy, Cambazoglu, and Bonchi, 2010),(McAuley and Leskovec, 2012). They based their classification on profile features such as common school, home town, education history, family name, group membership etc. The connections on OSNs may include friends, family members, workplace colleagues, acquaintances, etc. Many online social networking sites allow users to divide their connections into groups (also known as circles, lists, etc.) which is either accomplished by manual assignment or via automated assignment in a näive fashion. Both these approaches suffer drawbacks. The former may be too time consuming and the later may function poorly when profile information is missing or withheld.

In the subsequent sections we will:

1. Study composition of social circles with reference to attributes of nodes.
2. Analyze the interplay between interaction and social circles.
3. Infer friendship ties using interaction information.

6.2. Social Relations and Attributes

Attributes are qualities that describe an object. In social research, attributes refer to the characteristics of people or things (Babbie, 2010). The effects of social influence and social selection suggest that network ties and node attribute information should reiterate similar information. Gong et al. (2014) observed that attribute inference helped improve link prediction on a Google+ dataset.

In late 2014, Kaggle organized a competition called “Learning Social Circles in Networks”. Kaggle is a data science competition site. Given a dataset, participants are required to come up with a model that could extract certain information from the data. Scores are calculated by matching the results of participants with ground truth information. The task in this competition was to correctly identify social circles given the personal network and attribute information about the alters. The dataset, hand labeled by the egos, was referred to as the ‘ground-truth’. On this dataset, we analyze the characteristics of social circles with respect to different attributes, in order to see the variation in how egos identified their circles, as well as variation in the size and density of their personal networks.

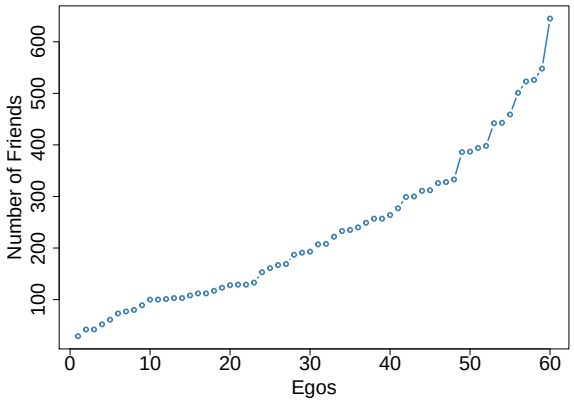
Data Description

Egonets: Each file corresponding to an ego contains the ego-network of a single Facebook user, i.e., a list of connections between their friends.

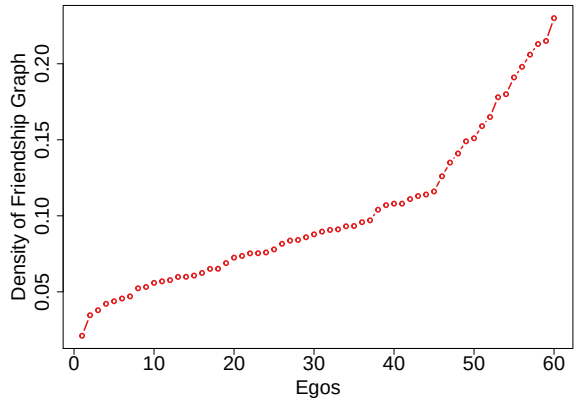
Attributes: Contains attributes for all users (egos and alters).

Training: Each file in the training set contains human-labeled circles provided by a user.

The dataset only contains data about the friendship ties and attributes and is devoid of any interaction data. Hence, in this dataset we can only study composition of social circles in terms of node attributes.



(a) Numer of friends. $\mu = 243$, $\sigma = 148$



(b) Density of Mutual Friendship Graph. $\mu = 0.10$, $\sigma = 0.05$

Figure 6.1.

Data statistics

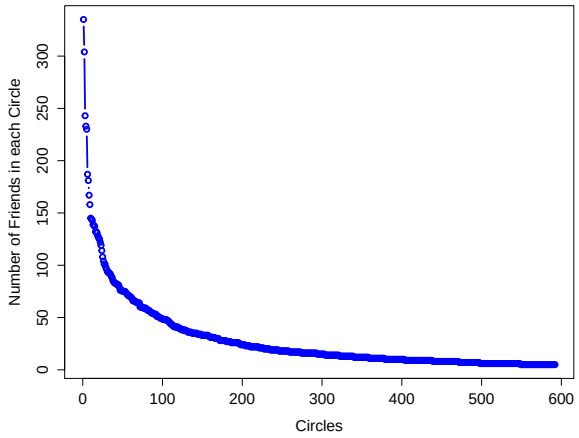
Distribution of friends: The *density of a graph* is the ratio of the number of edges to the number of possible edges. The average number of friends for each ego was 231. The distribution of friends was normal and positively-skewed, determined by the Kolmogorov-Smirnov test (KS test), which is a nonparametric test of the equality of continuous, one-dimensional probability distributions. It can be used to compare a sample with a reference probability distribution. Figure 6.1a shows the number of friends for each ego, whereas, Figure 6.1b shows how dense each ego network is.

Distribution of alters in each circle: We analyze the *number of alters in each circle* for each ego. Figure 6.2a shows the distribution of number of alters in all circles. There were 592 circles in all. Average number of alters in each circle was 27. Figure 6.2b shows the distribution of total number of circles for each ego. On average there were 9.8 circles per ego.

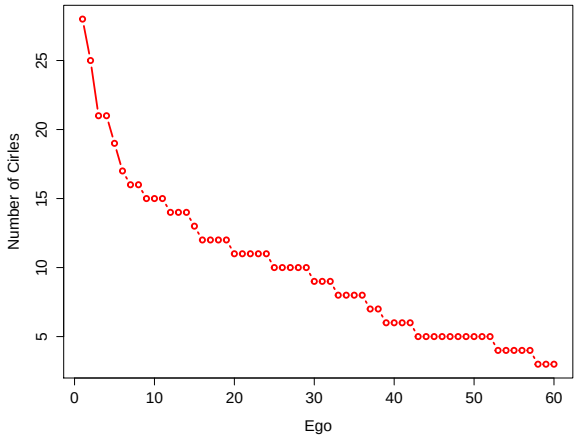
In order to study the variation in how circles are identified by egos, we compute the number of homophilous dyads in each circles and the homogeneity of each circle, based on the attributes of alters in the circle. We first define homophily and homogeneity:

Definition Homophily: A dyad is considered to be homophilous with respect to an attribute z, if the value of z is same for the two nodes sharing the dyad and there exists a tie between those nodes, whereas, a circle is completely homogeneous with respect to an attribute z, if the value of z is same for all the nodes in that circle.

We use the following attributes to evaluate the variation in composition of social circles: *last name, hometown, educational school, work employer, work projects id, location id, gender, worked with id, education-classes with, studied with, work on projects with id.* Some people meticulously placed every friend into some circle, with all circles completely disjoint, while some had just a few large circles with significant (over 80%) overlap. This indicates that the overall dataset was actually pretty small relative to the degree of variation in the data.



(a)

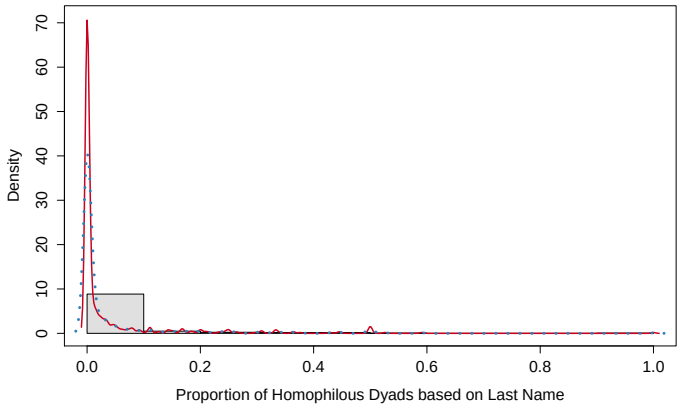


(b)

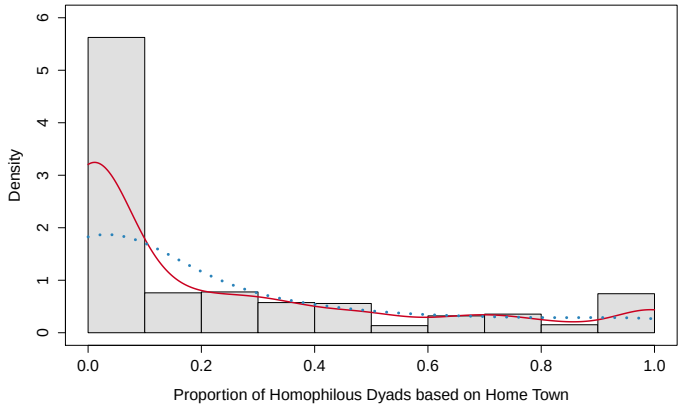
Figure 6.2.

Homophily: We computed the proportion of homophilous dyads for nine attributes. We then plotted the results in Figures 6.3-6.5. It is noted that gender homophily is very high in the social circles. In some social location(current location) and home town homophily is also high.

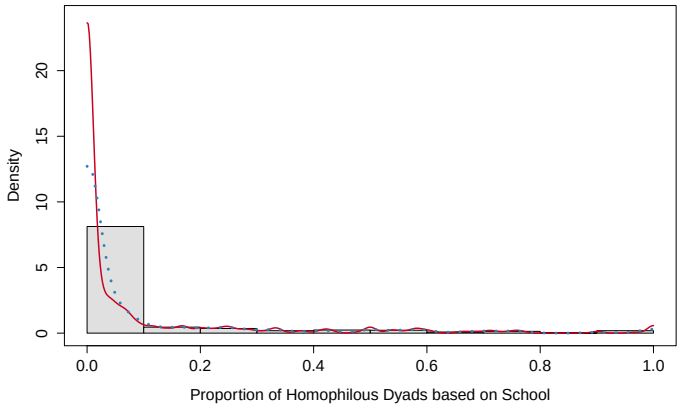
The first question we pose is whether circles contain significant homophilous dyads in any dimension? Figure 6.6 shows the percentage of homophilous dyads in every circle. For each circle we selected the dimension in which the homophily was highest. This was to show that circles are not randomly formed, rather each circle has a significant number of homophilous dyads in at least one dimension. As we see in the histogram, about 10% of circles had less than 40% homophilous dyads. About 25% circles had 100% homophilous dyads.



(a)

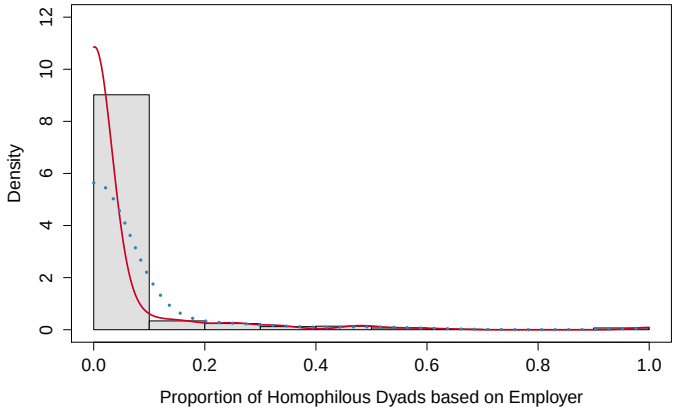


(b)

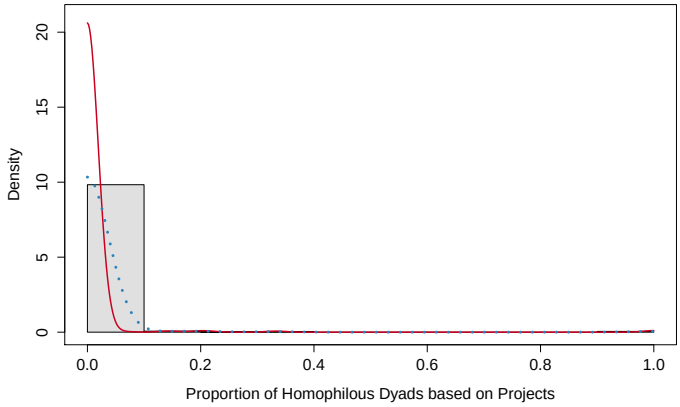


(c)

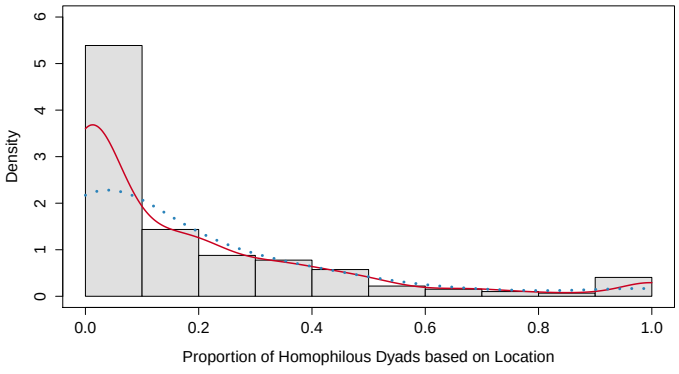
Figure 6.3.: Proportion of homophilous dyads based on different attributes



(a)

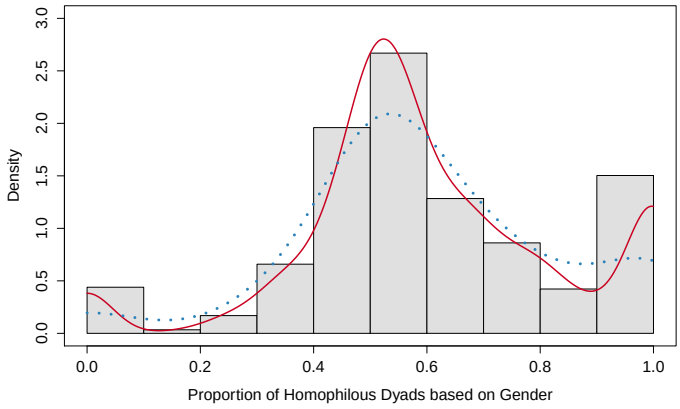


(b)

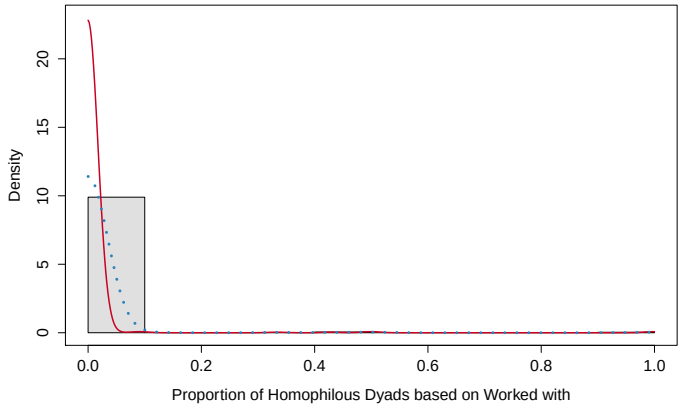


(c)

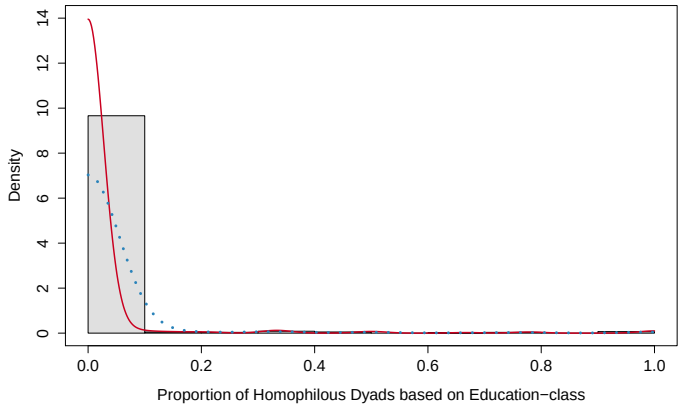
Figure 6.4.: Proportion of homophilous dyads based on different attributes



(a)



(b)



(c)

Figure 6.5.: Proportion of homophilous dyads based on different attributes

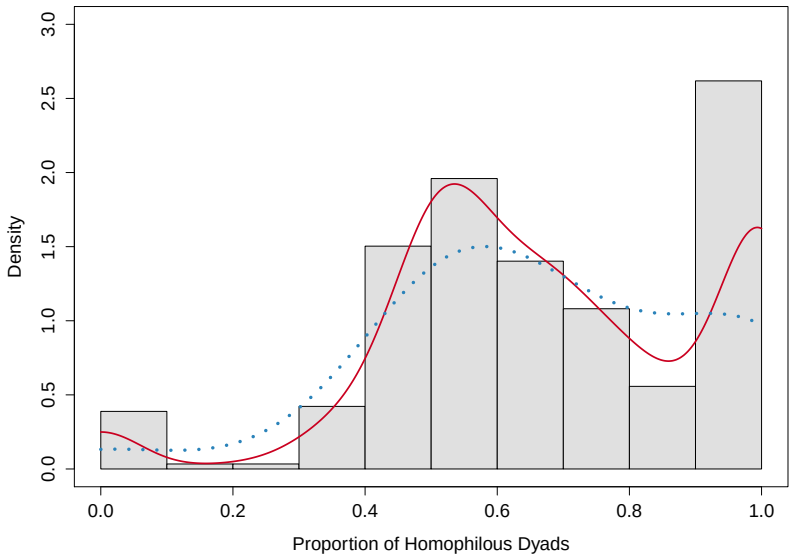


Figure 6.6.: Proportion of homophilous dyads in at least one dimension

We make the following inference about characteristics of social circles based on the analysis provided above. We also observed that the dataset was pretty small relative to the degree of variation in the data.

1. Social circles are homophilic in at least one dimension (attribute).
2. Gender homophily is very high in the social circles.
3. Location and Home town homophily is also significant in some social circles.

6.3. Interaction: A Representative of Social Circles?

In a preliminary experiment we first analyze the commenting behavior (Reusing the dataset from Nasim, Ilyas, et al. (2013)) using clique percolation method which is an overlapping community detection algorithm (Figure 6.7). Number of comments (y) on a post is shown on x-axis. Probability of receiving the next comment from the same community is plotted as $p_X[X = 1|N = n, Y = y]$ on Y axis. The probability of receiving the next comment from the same community is increasing given the number of comments from the same community are increasing. Figure 6.7 shows the results for $N = 1$ and $Y = 0, Y = 1$. This indeed shows that even for overlapping communities, the commenting probability increases when previous comments are from the members of the same community.

Problem Statement: Given commenting history of alters, can one infer social circles from discussions given the assumption that discussions take place within or between members of the same social circle. This assumption is motivated by Peter Blau's observation:

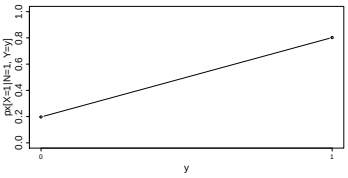


Figure 6.7.: Analysis of commenting behavior using Clique percolation which is an overlapping community detection algorithm.

> *"When people's differences in various dimensions are strongly correlated and consolidate social positions and group boundaries, people tend to refrain from intergroup relations even in those respects in which they have no ingroup bias (Peter M Blau, Beeker, and Fitzpatrick, 1984).*

In most of the previous work that was discussed in Section 6.1, focus was on clustering of mutual friendship graph (social relationship) or on clustering of discussions or mutual event participation(social interaction). In this section we would like to study discussions(interactions) proxy for foci and then use the mutual friendship (backbone relations) graph for validation.

We next visualize the social circles using the two-step process mentioned below and explain the findings and shortcomings using an illustrative example.

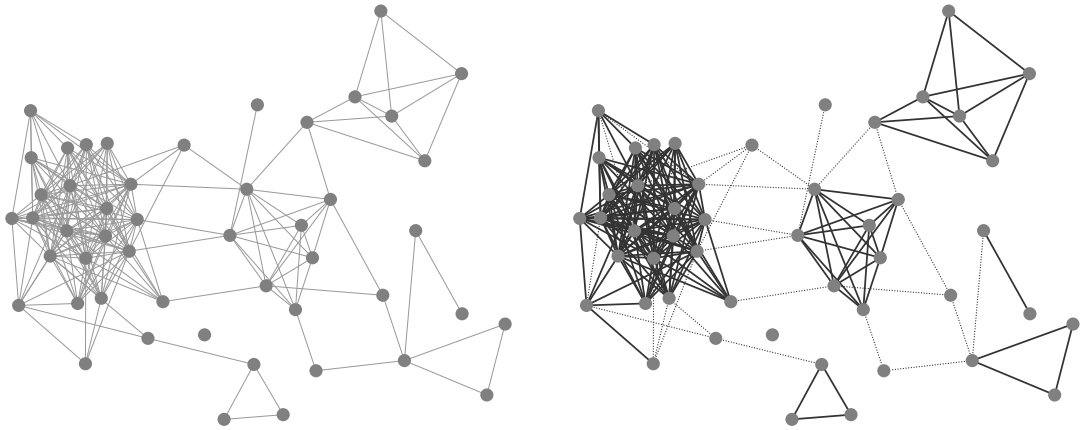
Filter for strongly embedded ties

In the first step we filter for strongly embedded ties using Simmelean backbone Nick et al., 2013. Simmelean backbone is motivated by Simmel's concept of membership in social groups. The method is capable of extracting core structure from dense networks such as Facebook networks, that makes them easy to visualize and analyze. The method uses a global valuation step, e.g. counting triangles or quadrangles (Nocaj, Ortmann, and Brandes, 2014) an edge is embedded in filtering based on local ranking.

The method performs as follows:

1. If input tie strength is uniform, assign to ties, the number of triangles they are embedded in (Simmelian strength).
2. For each actor, rank all alters by tie strength.
3. For each (strong) tie, determine its redundancy.
4. Filter ties that are weak or not redundant.

Figure 6.8 shows the graph transformation after applying Simmelean Backbone on a mutual friendship graph. We will use the same example network later in the text to map interactions on the core network.



(a)A mutual friendship graph

(b)Graph transformation after applying Simmelean Backbone

Figure 6.8.: An example network

Map interactions on core network

We now suggest a simple method to identify social circles of users based on the discussion threads they have participated in and see how well aligned network structure and social circles are with each other.

For a given ego, let $G = (V, E)$ denote the undirected personal network of an ego, where V is the set of alters and E is the set of friendship ties between the alters. Let $G' = (V, D, M)$ be a bipartite multigraph where V is the set of alters(users), D the set of discussions and $(u_i, d_j) \in M$ if user $u_i \in V$ has commented on discussion $d_j \in D$. Let A be the adjacency matrix of G'.

The entry at index ij is the number of comments by an alter $u_i \in V$ on post $d_j \in D$.

An ego can focus his friendship ties around many foci. Our goal is to study how the actions of alters reflect their social circles. For this purpose we cluster the posts based on the alters who comment together frequently. In this way, alters can be a past of more than one circle. A dissimilarity matrix is computed by using the distance measure to compute the dissimilarity between the rows of the matrix (discussions) A^T.

Classical multidimensional scaling (also known as principal coordinates analysis (PCA)) is then applied on the distance matrix. PCA takes a set of dissimilarities and returns a set of points such that the distances between the points are approximately equal to the dissimilarities. It is necessary to reduce these dimensions for clustering the posts.

Next we pick the top e dimensions from the resulting projection matrix. The top e dimensions is a $m \times d$ matrix. When the distance matrix is not Euclidean (which is true in our case), the projection matrix is not positive-definite. In such case, some of the eigenvalues are negative; correspondingly, some coordinate values have complex numbers. In order to pick the top e

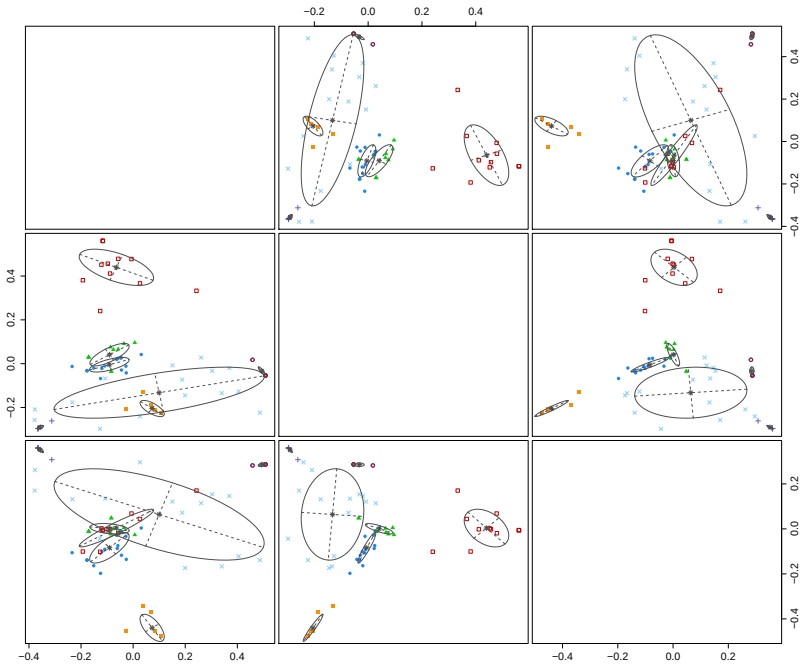


Figure 6.9.: Clusters of posts using the first three Eigen values

dimensions we determine the most significant eigenvalues $(\lambda_1, ..., \lambda_d)$, and define a threshold ratio $\theta \approx 0.80$ where,

$$\theta = (\sum_{i=1}^{e} \lambda_i^2) / (\sum_{i=1}^{m} \lambda_i^2) \tag{6.1}$$

Clustering of posts

Posts are clustered using the *Mclust*(Model based clustering)package in R (Fraley and Raftery, 2002). Mclust clusters, the data into optimal clusters according to Bayesian Information Criterion(BIC) for expectation maximization(EM) initialized by hierarchical clustering for parameterized Gaussian mixture models. The clusters generated by Mclust are centered at means μ_k. Clusters can vary in their shape. For multivariate analysis they can either be ellipsoidal, spherical or diagonal. One of the limitations of model based clustering is the number of parameters per component in multivariate normal mixtures that allow orientation to vary between clusters grows as the square of dimensions of data. Moreover, if the ratio of dimensions to the observations is higher, then the covariance estimates in the ellipsoidal models (which is mostly the case in our dataset) will often be singular, causing the EM algorithm to fail. We combine multidimensional scaling with Mclust considering the most significant dimensions.

Mclust clusters the posts and also outputs the mean of each cluster. However, some posts lie outside the boundary of the cluster. An example is shown in Figure 6.9. In such a situation we

would like to discard posts which are too far away from the center of the cluster. For that matter we calculate the distance d^M between the center of the cluster $C_k \subset C$ and the posts in that cluster in the multidimensional space. We calculate Mahalanobis distance for this purpose. The Mahalanobis distance provides a relative measure of a data point's distance from a common point. It differs from Euclidean distance because it takes into account the correlations of the data set and is scale-invariant. The Mahalanobis distance takes the coordinates of the posts in a given cluster and the covariance matrix(SC_k). Sometimes there are posts in the dataset which receive comments from the same alters which results in a singular covariance matrix, hence for the computations we use pseudoinverse(SC_k^+) of the covariance matrix instead of the original covariance matrix. Once the Mahalanobis distance is calculated, the decision to keep the post in a cluster is decided as a function of mean distance of all the posts from the center and the standard error.

If the Mahalanobis distance between the center of the cluster and the post is greater than mean plus the standard error, the post is discarded. Another choice we considered was to use the post with the minimum distance between the post and cluster center as the baseline, but that resulted in discarding a greater number of posts. The unique alters in the posts are identified for each cluster which we can now call a 'circle'.

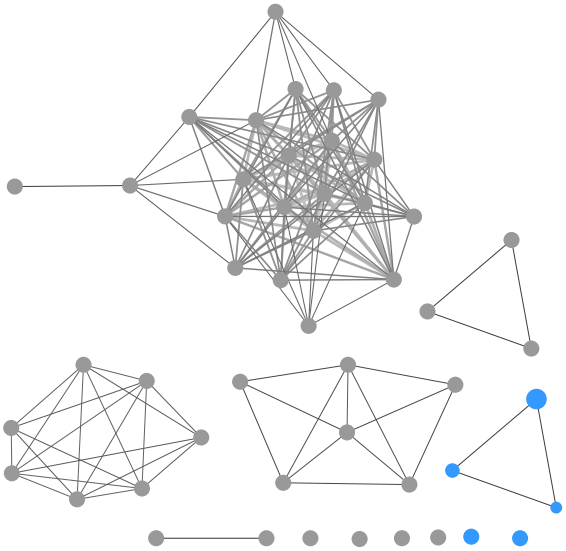


Figure 6.10.: Circle obtained from interaction information. Circle has a single focus

Figure 6.10 shows the ego network profile after removing the redundant links using the Simmelian backbone. The dark blue nodes show the one of the user's circles obtained from discussions. This circle has users that embedded in the core network and connected to each other. Figure 6.11 shows another circle that is not aligned with the network and has multiple foci. An interesting observation in Figure 6.10 is the presence of two blue nodes as isolates. Upon getting the hand-labeled data for this profile we found out that those isolates were in fact connected to the clique with all blue nodes. This further reinforces that circle obtained from the discussions have a single focus in Figure 6.11.

Although a high percentage of social circles is homophilic in at least one attribute dimension, the understanding, of what constitute a social circle is highly subjective as discussed in section 6.2. We were interested in finding accurate picture of social circles by aligning data about friendships and data about behavior i.e. commenting in our case and the motivation is that the mutual friendship graph is not an accurate representation of social circles, neither are discussions. By aligning discussions with the simmelian backbone we are identifying a strong signal. The strong embeddedness and discussions taking part in dense part of the backbone seem to have a clear form. We now shift our focus from the aggregate level study to dyadic level study where we can objectively determine the presence or absence of a friendship link.

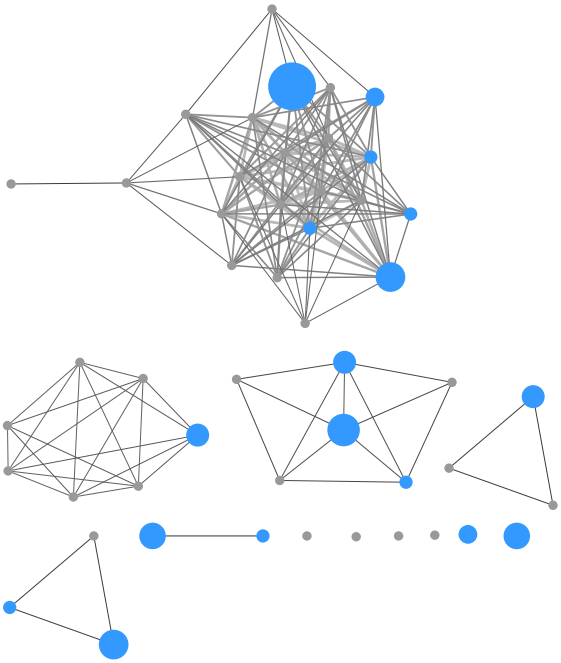


Figure 6.11.: Circle obtained from interaction information. Circle is not aligned with the network and has multiple foci. Size of the nodes corresponds to frequency of commenting, whereas, thickness of the edges indicates frequency of mutual commenting.

6.4. Link Inference

Sociological research has identified various dimensions of social relations, e.g., time, affect, intimacy, or reciprocal services (M. Granovetter, 1973) and group formation. We are interested in the question whether interactions in online social networks (OSNs) can serve as a proxy for more persistent social relation. With Facebook as the context of our analysis, we look at commenting on wall posts as a form of interaction, and friendship ties as social relations. In a formative study we investigate whether interactions can provide information on network ties even without content knowledge, i.e., instead of topical groups as defined, say, by hashtag usage, we make use of participation in discussions only, irrespective of their focus.

For a plausibility test of the hypothesis that interactions are closely related to ties, we use data originally collected to study commenting behavior of Facebook users (Nasim, Ilyas, et al., 2013). No topical or otherwise identifying information was recorded for status posts. The only information available is the identity of posters and commenters. For the rest of this section we collectively refer to a status post and comments on that post as a discussion.

Our goal is to infer friendship ties from participation in discussions. Suppose that we have the discussions related to a Facebook ego profile available to us but we do not have access to any information about the network structure, i.e., no friendship ties between alters are known

to us.

We will use the following notation:

- $u_1, ..., u_n$ denotes the n alters in an ego's personal network.
- $d_1, ..., d_m$ denotes the m discussions in a profile.
- $D(u_i)$ denotes the set of discussions in which u_i made a comment.
- $U(d_j)$ denotes the set of users who commented in discussion d_j.

Very simple features are extracted from the discussions to predict the presence of friendship relation between pairs of alters to determine the behavioral similarity of two alters based on the discussions they are part of. Recall that the discussions are not labeled unlike hashtags in Twitter where each hashtag defines a topic of the conversation. While, for example, *#socnet2014* may be a hashtag for discussion on SOCNET 2014, an ego may make several posts related to SOCNET 2014 on his or her Facebook wall leading to several distinct discussions. This implies that we are underutilizing the information available in principle.

Only the following features of discussions are used:

- $|D(u_i) \cap D(u_j)|$,
  the number of common joint participations of u_i and u_j.
- $min_{d \in D(u_i) \cap D(u_j)} |U(d)|$,
  the smallest size of a discussion group containing u_i and u_j.
- $max_{d \in D(u_i) \cap D(u_j)} |U(d)|$,
  the largest size of a discussion group containing u_i and u_j.
- $\frac{|D(u_i) \cap D(u_j)|}{|D(u_i) \cup D(u_j)|}$,
  the similarity of participation in discussion (Jaccard coefficient).

From these features we would like to predict the existence of friendship ties. Note that the data set does contain the actual friendship ties among alters of each ego, except for a few missing ones due to privacy settings. We thus picked one profile (shown in Fig. 6.12) for which the missing links could be added based on an interview with ego, and trained a simple regression model. Only pairs of users who have commented on the same post at least once are considered, and since there is a high degree of imbalance (with many more absent friendship ties than present ones) we randomly selected an equal number of adjacent and non-adjacent dyads for training.

We use logistic regression as our classification model. In logistic regression, the conditional distribution $y \mid x$ is a Bernoulli distribution because the dependent variable i.e. the class label(y) is binary. The outcomes Y_i are described as being Bernoulli-distributed data. p_i that is the unobserved probability determines the outcomes and it is related to the explanatory variables. It can be expressed as follows:

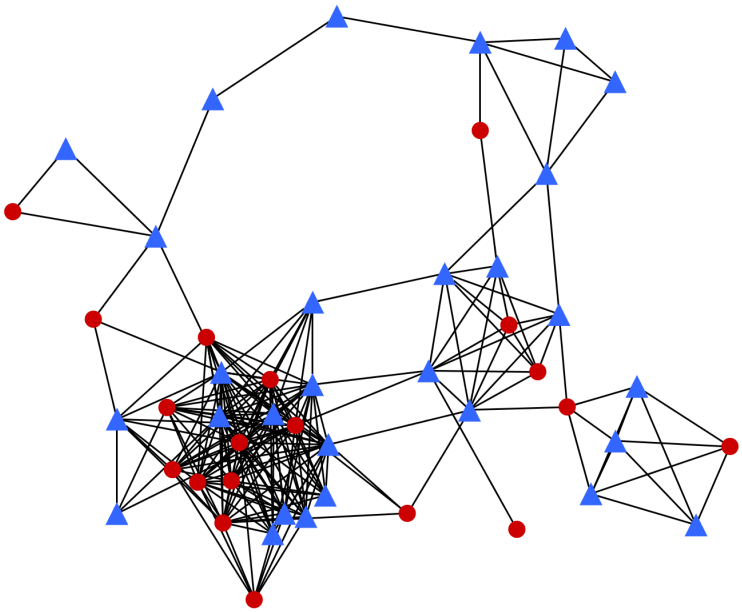


Figure 6.12.: Personal network comprised by the Facebook friends of a profile and their friendship ties. Alters who took part in discussions are shown as blue triangles. Note that the focal profile (i.e., the ego) would be connected to all others and is therefore omitted.

$$P(Y_i = y|x_{1,i},...,x_{m,i}) = p_i^y(i-p_i)^{(1-y)} \tag{6.2}$$

The model predicts probabilities through the logistic distribution. Logistic regression models the probability π using a linear predictor function, i.e. a linear combination of the explanatory variables and a set of regression coefficients that are specific to the model at hand but the same for all trials. Function $f(i)$ which is the linear predictor for data point i is written as:

$$f(i) = \beta_0 + \beta_1 x_{1,i} + ... + \beta_m x_{m,i} \tag{6.3}$$

where $\beta_0,...,\beta_m$ are regression coefficients. They indicate the relative effect of an explanatory variable on the outcome.

For comparison, we also trained the model from Horvát et al., 2012, in which structural features computed on an induced subgraph (representing members of an OSN) are used to predict links between the other nodes (the non-members).

Evaluation and Results

We used 'accuracy' to test the performance of our classifier. This is a statistical measure of how well a binary classifier correctly identifies/excludes a condition. It is defined in terms of True Positives (tp), False (fp), True Negatives (tn) and False Negatives (fn).

$$accuracy = \frac{tp + tn}{tp + fp + tn + fn} \tag{6.4}$$

We motivate the features and explain the results on the example shown in Fig. 6.12. It has been used for training the model once as it was recorded and once with the missing links introduced with the help of ego. The results are reported in Tabs. 6.1 and 6.2.

Table 6.1.: Prediction accuracies after training on recorded data

Features	Min accuracy	Avg. accuracy	Max. accuracy
Number of common discussions	36.68%	55.92%	72.58%
Size of smallest discussion group	40.06%	54.79%	73.75%
Size of largest discussion group	34.29%	52.63%	70.08%
Jaccard coefficient	35.25%	53.02%	78.31%
All features	49.67%	62.17%	80.62%

Table 6.2.: Prediction accuracies after training on interview-corrected data

Features	Min. accuracy	Avg. accuracy	Max. accuracy
Number of common discussions	49.12 %	58.30%	79.69%
Size of smallest discussion group	47.92%	58.74%	74.01%
Size of largest discussion group	46.57%	56.38%	71.56%
Jaccard coefficient	57.91%	61.45%	84.53%
All features	54.61%	68.92%	82.23%

One intuitive way to measure the similarity between alters is to find the number of common discussions they participated in. However, this measure does not differentiate between the discussions which are of broad interest and those which are of interest to people who share a common foci. The classification based on this feature gives an accuracy of 55.92%. Another way to measure the exclusiveness of the discussions is to consider the minimum/maximum number of participants in a discussion. Nevertheless, these features do not differentiate between alters who always share exclusive discussions versus the alters who less frequently participate in exclusive discussions. We also use Jaccard coefficient as a discussion feature. The prediction accuracy with Jaccard coefficient is anticipated to be higher, however the table shows a lower prediction accuracy on average. When we added the missing links in the data for training the algorithm, the accuracy of Jaccard coefficient significantly improved. The combined accuracy of discussions' features is 62.17%. The prediction accuracy shows both the correct number of 1s and 0s predicted. The p-values we obtained for the logistic regression show that the first

three features are statistically significant. The estimate coefficient in the case of number of common discussions is positive, which means higher the number of common discussions, the more is the probability that two individuals have a friendship tie. The estimate coefficients for the size of smallest common discussion group and the size of largest common discussion group are negative. This signifies that smaller the size of a discussion group, the more exclusive it would be which is indicative of the fact that participants involved in the discussion share friendship ties.

As mentioned earlier the data set had some missing friendship links; we mitigated the problem of false negatives by filling in the missing links between alters in order to train our algorithm on correct data. The training profile had six missing links between alters. We then analyzed the amount of network information reflected in the discussions. The prediction accuracies are summarized in Tab. 6.2. Using all the discussion features, on average we get a prediction accuracy of 68.92% and in some cases it can be as good as 82.23% . Adding the information about the missing links in our training/testing set also shows an improvement in the prediction accuracy of the Jaccard coefficient feature. This feature signifies the similarity between two alters and shows how exclusively they comment together over their entire commenting history.

In order to compare our work with another state-of-the-art link prediction approach, we also implemented the algorithm of Horvát et al. (2012) which is based on structural properties of the network. The basic approach classifies nodes into two categories, social networking site members and non-members. In our experiments, the selection of members is based on independent decisions modeled by the random selection of a set of members; the value of parameter ρ is 0.5 (i.e., 50% of the nodes are members) and α is 0.8 (i.e., 80% of the members have shared information about the edges). The structural features mentioned in the study work well when the features are constructed on erroneous data (i.e., data with few missing links), since the predicted variable may also have several missing links. In our experiments we constructed the structural features on erroneous data, but the predicted variable was inferred from the correct data. Our experiments show that approach used by Horvát et al. (2012), gives a prediction accuracy of 65.21%. We then used the interaction features along with the structural features and we get a prediction accuracy of 80.43%. Features based on structural properties of graph are incapable of capturing links where not only the two nodes under consideration have their friendship lists hidden but also the friendship lists of their neighbors are hidden. In such scenarios interaction information acts as a proxy to friendship links.

6.5. Discussion

In this chapter we analyzed datasets from online social networks to study whether instantaneous relations (interaction) can point to more persistent relations (friendship ties).

Group membership is equally important and essential for human survival as dyadic relationships are. We found that based on users' perception of friendship groups in their online social networks, *social circles*, are homophilic in at least one dimension (attribute).

M. Goldberg et al. (2010) and Romero, Tan, and Ugander (2013) also studied interaction information and community structure on blogs and twitter respectively. There are three crucial difference between our work and M. Goldberg et al. (2010) . We are looking at the social circles from the perspective of an ego. Goldberg et al. presented a general structural and attribute based verification showing that overlapping communities frequently occur in social networks. In our analysis we have limited knowledge about the network i.e., only the personal networks of distinct egos are available. In a hypothetical situation, a node pair may not share any neighbors in a given personal network but in the original network they could be part of a large clique. This is something which cannot easily be captured by structural approaches. Goldberg et al. made use of one to one communication between the users whereas we are looking at joint commenting of alters on an ego's wall. The third crucial difference we find in the scheme of Goldberg was that blogs with single, dual or multiple focus were treated equally. In our work we first distinguish discussions based on their focus by looking at the discussion features and their relationship with the network structure. As a first step we showed that by aligning discussions with the simmelian backbone we are identifying a strong signal. The strong embeddedness and discussions taking part in dense part of the backbone seem to have a clear form. Since the understanding of social circles is a subjective to users' perception we then moved on to establishing a relation between commenting information and social relations given that the information about friendship ties is readily available.

Romero, Tan, and Ugander (2013), used features of common hashtags to predict links between Twitter users. They considered the follow edges and @edges and used the common hashtags and size of minimum/maximum common hashtags as their features, in addition to using the aggregated features of hashtags. In contrast to using discussions sizes as the aggregate overlap of discussions, we find it prudent to use the Jaccard coefficient to measure the similarity between interaction history of pairs of alters in the context of Facebook. For the follow edges the prediction accuracy of Romero, Tan, and Ugander (2013) is 74% as compared to the 68% accuracy we are getting for the undirected friendship edges on Facebook. Recall, however, that our discussions are not classified into topics, unlike the hashtags in Twitter. This makes the prediction task more difficult. The @edges require knowledge about the content of the tweets and in our case the content of the discussions. Moreover, they also show a boost in prediction

accuracy when information about network edges (other than the ones connected to the two users being considered) is introduced.

We were interested in the question whether interactions in online social networks (OSNs) can serve as a proxy for more persistent social relation. With Facebook as the context of our analysis, we looked at commenting on wall posts as a form of interaction, and friendship ties as social relations. Findings from our pretest suggest that others' joint commenting patterns on someone's status posts are indeed indicative of friendship ties between them, independent of the contents. This would have implications for the effectiveness of privacy settings.

In the next chapter we will look at an application (an in depth case study) of our findings.

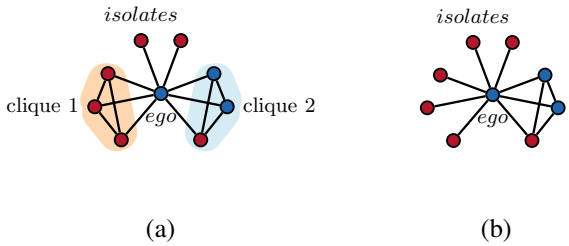
7. Applications: Interaction as a Proxy for Network Structure

7.1. Link Inference in Partially Observable Online Social Networks

Simmel's theory of social circles posits that a person has several personality aspects owing to his or her membership in different social groups (Simmel, 1908). These include both real life social groups (people are friends with workplace colleagues, school mates, family members, organization members, etc. (Feld, 1981)) as well as social groups on online social networking sites (Mcauley and Leskovec, 2014). The social roles tend to reinforce each other. There may be pressure to conform to those social roles. Peter Michael Blau (1977) conceived social structure as a space in which positions are determined by the individuals' characteristics. Homophily, then, is the assumption that individuals farther apart in this space are less likely to interact with each other. It is argued that social constraints may impede intergroup relations which means that when homophily is correlated along multiple social constraints within communities, it can lead to limited intergroup interaction.

These theories thus suggest that interactions are likely to take place among individuals who are also connected. In fact, repeated interaction is often conceived as a prerequisite for the establishment of social ties. Hence we attempt to discover unobserved links in online social networks with the help of interaction via participation in joined discussions, but without considering the content of these discussions. Taking a small personal network as an example, we illustrate the problem at hand in Figure 7.1. Nodes in blue are the ones whose network information is available (e.g., visible circles membership in Google+ or friends on Facebook). Nodes in red are the ones whose network information is not visible. In the original network, a link between a blue and red node can be inferred by the information acquired from the blue node (assuming there are indirect links). However, the situation becomes challenging when one has to infer links between any two red nodes, i.e., where the network information is not visible.

Studies on link prediction have focused on properties such as existing network structure, actor attributes and interaction patterns to deduce information about the users. In the pioneering



(a) (b)

Figure 7.1.: Personal network example. (a) actual network, (b) observed network

work on link prediction, Liben-Nowell and Jon Kleinberg (2007) considered a set of features limited to the topological properties of the network. The approach is generic and can be applied to any social network graph. Following these foundations(Liben-Nowell and Jon Kleinberg, 2007), various network topology based approaches have been devised to predict links in social networks (Adamic and Adar, 2003)(Leskovec, Huttenlocher, and Jon Kleinberg, 2010)(Horvát et al., 2012). A major limitation of topology based features is observed when the network information is significantly missing which may lead to erroneous training set and eventually affect the performance of the classifier.

In this work we show that interaction information can help infer unobserved (e.g. missing or hidden) relational ties more accurately. The successful inference can help uncover the links in networks which otherwise are not observable. For example, inferring friendship links on online social networking sites where friendship information can be made hidden (e.g. Facebook, Google+, Flickr), or inferring links from information about event participation, group discussions, etc. The accurate prediction of links in the presence of minimal network information could help discover the relation between communication and connectivity patterns in online social networks. On the one hand identification of important missing links can be used for product and social recommendation systems, on the other hand inference of network information poses a serious threat to the privacy mechanisms incorporated in social networking sites.

7.2. Conventional Approaches to Link Prediction

Studies on link prediction have focused on properties such as existing network structure, actor attributes and interaction patterns to deduce information about the users (Al Hasan and Zaki, 2011). In this section we will discuss various link prediction methods found in the literature.

Network Structure: Liben-Nowell and Kleinberg's (Liben-Nowell and Jon Kleinberg, 2007) seminal work on link prediction problem considered a set of network features. They

computed the similarity scores between nodes based on different measures which depend on node neighborhoods (e.g. using Jaccard coefficient, Preferential attachment, etc.), ensemble of all paths (e.g. Katz coefficient, PageRank, SimRank, etc.) and higher level approaches (Low rank approximation, Unseen bigrams and Clustering). The best predictor i.e. Katz coefficient was correct only about 16% of the predictions. Taking this work a step further, Cukierski, Hamner, and B. Yang (2011) looked into the accuracy of graph-based features for supervised link prediction. They tested multiple network based features on a dataset from Flickr. They found that the best classification results for a supervised link prediction problem were achieved through a combination of a large number of features which could capture various aspects of the graph structure. Horvát et al. (2012) showed that the combination of knowledge of confirmed contacts between members on a social network and their email contacts to non-members provides enough information to deduce a substantial proportion of relationships between non-members. They formulated the link prediction problem as a binary classification problem. They used topological features to predict links on five different Facebook datasets. Since the ground truth information about the non-members was not visible, therefore they emulated the non-members by labeling a fraction of Facebook friends as non-members in their datasets. Leroy, Cambazoglu, and Bonchi (2010) used interest groups on Flickr to infer hidden links. In a similar study (Zheleva et al., 2010) authors showed that when there are tightly-knit family circles, the accuracy of link prediction models can be improved. They made use of the family circle features based on the structural equivalence of family members. An important category features they used were group features. Those are the features that overlay friendship and affiliation networks. The features for all node pairs include number of friends in the family and portion of friends in the family. They also based their assumption on homophily. They assumed that nodes in each group are likely to behave similarly. New links can be predicted by projecting links such that the nodes in the group become structurally equivalent. However, in order to define relationship between an actor and a group of other actors, friendship (network) information is used.

Attributes: Both social influence and social selection suggest that network structure and node attribute information should be reinforcing concepts. Several studies have therefore looked into improving link prediction with attribute information (Gong et al., 2014)(Yin et al., 2010). Mislove, Viswanath, et al. (2010) inferred attributes of users, in combination with the social network graph on the premise that groups in the network are formed around users who share certain attributes. Link-prediction methods utilizing attribute information first appeared in the relational learning community. Taskar et al. (2003) addressed the problem of link prediction in relational domains. They have focused on the task of collective link classification, where they simultaneously predict and classify an entire set of links in a link graph. They use topological

properties and attributes of nodes to define a single probabilistic model over the entire link graph. Modeling link graphs has numerous other applications, including: analyzing communities of people, identifying people who may play certain key roles and also predicting current and future interactions. S.-H. Yang et al. (2011) proposed jointly predicting links and propagating node interests (e.g., music interest). They showed that the interest and friendship information is highly relevant for suggesting friends in social networks.

Interaction: Event-based network data consists of sets of events over a period of time. Examples of such networks are email traffic, co-authorship events, telephone calls, etc. O'Madadhain, Hutchins, and Smyth (2005) looked at the problem of temporal link prediction for co- authorship and email networks. They argued that using techniques from data mining and machine learning can yield scalable robust algorithms for predictive modeling. Temporal interactions from call-logs have been used to predict links and community structure in social networks (Eagle, A. S. Pentland, and Lazer, 2009)(M. Goldberg et al., 2010)(Tabourier, Libert, and Lambiotte, 2016). Lee et al. (2013) demonstrated (using social vector clocks) that link predictors can be made more effective if they operate directly on longitudinal dyadic communication data. Nasim, Ilyas, et al. (2013) provided an insight on the commenting behavior of Facebook users. Their results suggested that just like offline behavior, people tend to conform to group dynamics when they are on online social networking sites and their commenting behavior is a result of social influence. Romero, Tan, and Ugander (2013), studied the interplay between network and topical structure on Twitter. They inferred links between users on Twitter. The authors investigated the topical affiliation system defined by the usage of hashtags. The study uses hashtags to define user sets which can be viewed as Twitter users embedded in the topics associated with those hashtags. Users who followed same tweets were more likely to have links between them. One important thing missing in this study was the dynamics behind the follower and followee relations. They did not distinguish between these relations. For instance if a famous person posts a tweet, it is likely that her fans are going to retweet it using the same hashtags, hence the link between the follower and followee becomes easy to predict as compared to the case where two random twitter users use the same hashtag. The study also made use of '@' messages which on its own a reasonable predictor of follower/following relation on Twitter. The '@' sign is used to call out usernames in Tweets, e.g., Hello @Twitter! When a username is preceded by the '@' sign, it becomes a link to a Twitter profile (*Twitter Glossary* n.d.). In order to predict links on Facebook, Backstrom and Leskovec (2011) used edge creation time, edge initiator, probability of communication and profile observation in a one week period, and number of common friends as prediction features. Their method showed 11% relative improvement in Prec@20 as compared to Random Walk with Restart and Logistic Regression (using node+network features). An interesting finding in the paper is that the most

important features for Facebook are the ones related to time. The authors acknowledge that extracting good features that describe the network structure and connectivity patterns between the pair of nodes under consideration is a challenging task.

7.3. Inferring Links Using Interaction Information

In the first quarter of 2015, Facebook had 1.44 billion active users which rose to 1.65 billion active users in 2016 (Statista, 2016). This makes Facebook the most actively used social networking site. Facebook allows its users various interaction options such as sending messages, participating in events, sharing pictures, videos, status posts etc.

Issues related to privacy of shared content and personal information has received significant attention not only in the research community but also in the media. The privacy settings for content shared on Facebook match users' expectations only 37% of the time; if incorrect, then the content is usually exposed to more users than expected (Y. Liu et al., 2011). Facebook also provides privacy mechanisms to protect friendship information such as hiding friends from public or other contacts.

In case friends information is hidden, we propose that a malicious contact who has access to a user's 'timeline' may not only infer user's active friends but also the connections between them. Facebook, as well as third parties, can create applications and games which Facebook users can add to their profiles. These apps have access to users' profile data such as list of friends, wall posts, demographic information, etc Wang, Xu, and Grossklags, 2011. Due to privacy concerns related to the use of data obtained by third-party applications, Facebook restricted access to several data fields in its API. As of April 30, 2015, friend list now only returns friends who also use the same app/game. Our question is hence: Given these privacy mechanisms, is it still possible for a third-party app (or a bot) to determine the personal network of a user who is using that application?

We demonstrate that interaction information can boost the prediction of unobserved (missing or hidden) relations in partially observed networks. Information on ties in online social networking sites can thus be obtained even when it can be hidden or withheld. While such information can potentially be used for product and social recommendation systems, it undermines the privacy mechanisms users of such systems rely on.

Problem Statement: Formally, the link inference problem in a partially observable network can be described as follows:

Given an undirected graph (mutual friendship graph of an ego) $G = (V, E \cup E')$, consisting of the set of V vertices (alters in this case) and set of E edges (friendship links between alters), find the set E' of missing edges. We propose to design a binary classifier which uses set of

features (X) with cardinality $|X| = m$ extracted from given network and interaction pattern of alters in order to give accurate inference of links that are likely to be present in the graph G. The extraction of these features is described in the subsequent sections. The features are computed for all dyads. Input to the classifier are the network features ($X_1, ..., X_p \in X$) and interaction features ($X_{p+1}, ..., X_m \in X$). The k^{th} feature for node pair i and j is denoted by X_k^{ij}.

The outcome, i.e. the class label $(Y_{ij}|X_1^{ij}, ..., X_m^{ij})$ is Bernoulli distributed data conditioned on features $(X_1^{ij}, ..., X_m^{ij} \in X)$ and therefore $Y_{ij} = 1$ if there is a link between node pair i and j and $Y_{ij} = 0$ otherwise.

7.4. Case Study

Existing datasets in the public domain contain limited information in the context of this study. Most publicly available datasets (e.g. Facebook, Flickr and Google+) contain network and attributes information but are devoid of any interaction history or contain very limited information about the interaction between nodes in the network. The Stanford Network Analysis Project (SNAP) Leskovec and Krevl (2014) host many such datasets. Datasets that contain interaction information, for example the datasets used in Mondal et al., 2014, and Tang, S. Wu, and Sun, 2013, are not available. For this study we used an original Facebook dataset provided by the Algopol project [1] as part of a survey among volunteer participants selected by a poll institute to make a representative panel of French Facebook users (Bastard et al., 2015). A Facebook application collected data from Facebook profiles. Data was completely anonymized before analysis. The application downloaded the mutual friendship graph, wall posts and attributes of ego and his/her friends.

The dataset contains 586 ego profiles, where 38% of egos are male, 60% female and remaining 2% did not report the gender. There are 64,000 friends in total, of which 52% are females, 46% males and 2% did not report the gender.

We model the mutual friendship information collected from each ego profile as undirected graph $G = (V, E)$, consisting of the set of V vertices and set of E edges. An edge between two vertices indicates a friendship tie. Facebook provides various interaction options to users. One of them is posting on one's wall. Friends can then write comments on that post. Graphically, for each ego profile, we can represent the set of posts $\{p_1, .., p_n\} \in P$ and the commenters $\{c_1, ..., c_m\} \in C \subseteq V$, as a bipartite graph (a two-mode network, and the derived one-mode network of commenters), as shown in Figure 7.2.

[1] http://algopol.huma-num.fr/ grant #ANR-12-CORD-013

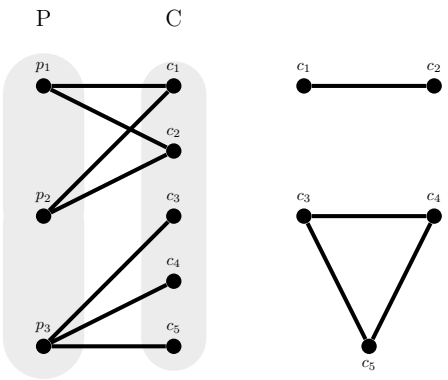


Figure 7.2.: A two-mode network showing posts and commenters (left). One-mode network of commenters is derived (right).

7.4.1. User Behavior Model

Facebook allows its users various interaction options, e.g. wall posts, commenting, liking posts, sending private messages etc. In this study we are looking at the structural features of comments received on wall posts of an ego. Content of the posts and comments is not analyzed.

Posts

We model the posting behavior of an ego by the nature of the posts. Our assumption is that either the topic of the post is one of global interests or of local interest. In the former case the post is likely to attract the interest of many of the ego's friends. Examples of such posts include: "*I have graduated*", "*Expecting an addition to our family :-)*", etc. In the latter case when the post is of limited interest, it is likely to attract interest of friends who are specifically interested in the topic. Examples of such posts include: workplace gossip, sports club related posts, family related posts, etc.

Commenting

We model the decision of a Facebook friend u by a random variable Z_u, where $Z_u = 1$ if u comments on the post and $Z_u = 0$ otherwise. Further, we are aware that the decision to comment depends on many factors such as: *Previous comments, user interest, decay-time, propensity to interact and how active the user is on the social networking site*, etc. These factors are catered for in the wide variety of features we are extracting from the data. An important assumption is that the future comments on a post are affected by prior comments on the post (Nasim, Ilyas, et al., 2013)(Tang, S. Wu, and Sun, 2013).

We call discussion a collection of comments on a single wall post.

7.4.2. Data Partitioning

We have formulated the link inference problem as a supervised binary classification problem. Supervised learning methods are apt at dealing with datasets which have greater class imbalance (e.g. online friendship networks) (Lichtenwalter, Lussier, and Chawla, 2010). In our study we look at friendship networks which are very sparse and thus have a high class imbalance.

We divide the data into training and test sets. The model is trained on the training sample and is used to classify the class labels in the test sample. We use a 10-fold cross-validation. For every fold we train the algorithm on 90% of the networks and test the performance on the remaining 10%. We have ensured mutual exclusivity between data points used in training and test samples in the experiments.

7.4.3. Feature Extraction

For predicting the presence of a link between two nodes u and v, we compute certain features pertinent to the network structure and discussions for all node pairs $(u,v) \in V$ where $u \neq v$. Description of features is mentioned in Table 7.1.

Network: We compute eight different network features, out of which six are local features. They are: common neighbors, common neighbors normalized by min/max degree of each node, Jaccard coefficient, Adamic Adar distance, Preferential attachment score and Cosine similarity. We also computed Katz score. All but one (Katz similarity) are local features based only on neighborhoods $N(u)$ and $N(v)$. Note that the number of common neighbors equals the number of triads closed when u and v are linked.

These features represent the state-of-the-art for predicting links in social networks in both theoretical (Sarkar, Chakrabarti, and Moore, 2011) and empirical studies (Aiello et al., 2012)(Backstrom and Leskovec, 2011)(Cukierski, Hamner, and B. Yang, 2011)(Tarissan, Latapy, and Prieur, 2009).

Discussions: Six network features are transferred to the interaction space, where $S(u)$ and $S(v)$ represent the discussions in which the alters participate. Note that we do not make use of labels and thus do not discriminate discussions by, say, topic.

7.4.4. Feature Selection

Filtering for features helps assessing the merits of attributes from the data while ignoring the learning algorithm. We used the following two feature ranking methods:

Table 7.1.: Features for predicting a tie between alters u and v.

Name	Formula	Description
Common neighbors	$\lvert N(u) \cap N(v) \rvert$	Size of intersection of the two neighborhoods
Jaccard similarity	$\frac{\lvert N(u) \cap N(v) \rvert}{\lvert N(u) \cup N(v) \rvert}$	Common neighbors normalized by joint neighborhood size
Common neighbors (min-normalized)	$\frac{\lvert N(u) \cap N(v) \rvert}{\min\{deg(u), deg(v)\}}$	Number of common neighbors normalized by smaller neighborhood
Common neighbors (max-normalized)	$\frac{\lvert N(u) \cap N(v) \rvert}{\max\{deg(u), deg(v)\}}$	Number of common neighbors normalized by larger neighborhood
Adamic-Adar	$\sum_{k \in N(u) \cap N(v)} \frac{1}{log(deg(k))}$	Similarity based on degrees of common neighbors
Preferential attachment	$\lvert N(u) \rvert \cdot \lvert N(v) \rvert$	Product of neighborhood sizes
Cosine similarity	$\frac{\lvert N(u) \cap N(v) \rvert}{\lvert N(u) \rvert \lvert N(v) \rvert}$	Common neighbors normalized by preferential attachment
Katz similarity	$\sum_{l=1}^{\infty} (\alpha A)^l$	Walks from u to v weighted by length
Joint discussions	$\lvert S(u) \cap S(v) \rvert$	Number of discussions both alters participated in
Joint discussions (min-normalized)	$\frac{\lvert S(u) \cap S(v) \rvert}{\min\{\lvert S(u) \rvert, \lvert S(v) \rvert\}}$	Joint discussions normalized by smaller participation set
Joint discussions (max-normalized)	$\frac{\lvert S(u) \cap S(v) \rvert}{\max\{\lvert S(u) \rvert, \lvert S(v) \rvert\}}$	Joint discussions normalized by larger participation set
Jaccard (discussions)	$\frac{\lvert S(u) \cap S(v) \rvert}{\lvert S(u) \cup S(v) \rvert}$	Joint discussions normalized by combined set of discussions
Smallest discussion size	$\min_{d \in S(u) \cap S(v)} \lvert U(d) \rvert$	Smallest number of participants in joint discussions
Largest discussion size	$\max_{d \in S(u) \cap S(v)} \lvert U(d) \rvert$	Largest number of participants in joint discussions

Chi-square statistic

Chi-square (χ^2) is the simplest of filters for feature filtering. It is defined as:

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected} \tag{7.1}$$

It measures divergence from the expected distribution if one assumes the feature occurrence is independent of the class value.

Table 7.2.: Feature ranking results using two filters: Information Gain (IG) and Chi-square (χ^2) for networks with $50-100\%$ edges removed. Please note that when 100% edges are missing, network features are not applicable.

	50%		75%		90%		100%	
Feature name	IG	χ^2	IG	χ^2	IG	χ^2	IG	χ^2
Katz	1	1	1	1	1	2	NA	NA
Cosine	2	2	2	4	3	3	NA	NA
Jaccard	3	3	5	5	10	4	NA	NA
Adamic Adar	4	4	3	2	11	5	NA	NA
Common neighbors	5	5	4	3	12	7	NA	NA
Preferential attachment	6	6	6	6	2	1	NA	NA
Overlap coefficient for common neighbors	13	13	13	13	13	13	NA	NA
Common neighbors normalized (max)	14	14	14	14	14	14	NA	NA
Smallest discussion size	7	7	7	7	4	6	1	1
Largest discussion size	8	9	8	8	6	9	3	3
Overlap coefficient for discussions	9	8	9	9	5	8	2	2
Common discussions	10	10	10	10	8	10	5	4
Jaccard discussions	11	11	11	11	7	11	4	5
Common discussions normalized	12	12	12	12	9	12	6	6

Information gain

Information gain is a a measure based on the information- theoretical concept of *entropy*, which is a measure of the uncertainty of a random variable. Entropy of a variable, (in our case a feature) X is defined as follows:

$$H(X) = -\sum_i P(x_i)\log_2(P(x_i)) \tag{7.2}$$

and the entropy of X conditioned on another variable (in our case class)Y is given by:

$$H(X|Y) = -\sum_j P(y_i)\sum_i P(x_i|y_j)\log_2(P(x_i|y_j)) \tag{7.3}$$

Amount by which the entropy of X decreases indicates additional information provided by Y and this is called information gain or mutual information Cover and Thomas, 2012. For a

feature X_i with respect to class label Y, information gain is defined as:

$$IG(X_i|Y) = H(X_i) - H(X_i|Y) \tag{7.4}$$

7.4.5. Classifier Design

We show our results using Logistic Regression (LR) model. Logistic regression measures the relationship between the categorical dependent variable and one or more independent variables, by estimating probabilities. Logistic regression is very scalable and does not assume that the features are normally distributed (unlike LDA). For a social networks data set with nearly 2 Million data points, LR is the best suited classification algorithm. In order to avoid limiting the analysis to one algorithm, we evaluated four supervised learning algorithms: Logistic Regression (LR), Linear Discriminant Analysis (LDA), Naive Bayes (NB) and Support Vector Machines (SVM) using the implementation available in R (R Core Team, 2014),(Meyer et al., 2015). Since support vector machines with non-linear kernels may take a very long time to execute, it is recommended to use SVMs with linear kernels when there are more than 50,000 data points in the training set. SVM with linear kernel is similar to logistic regression.

7.4.6. Performance Evaluation

Data Sampling: We employed two methods for data sampling:

1. In order to hide some links, we randomly remove edges from each network and then evaluate the performance of our classifier. We call those links hidden or missing links.. We have restricted the removal to four cases: I) 50% edges missing from the network, II) 75% edges missing from the network, III) 90% edges missing from the network and IV) there is no network information available. Please note that features are recomputed for each case after removing the edges (both in training and test sets).
2. We sampled our network in such a way that the links incident on certain nodes are removed. We removed edges incident to nodes until a certain percentage of nodes (V^*) are isolates. We have restricted $V^*$to $50\%, 75\%, 90\%$ and 100%. This is to create a situation similar to the one where a certain percentage of ego's friends have made the friends list hidden as discussed in the introduction section (Please refer to Figure 7.1). Please note that features are recomputed for each case after removing the edges (both in training and test sets).

We compare the performance of discussions features with some of the established network based link prediction features found in the literature (Cukierski, Hamner, and B. Yang, 2011).

Evaluation metric: There are different ways to measure the performance of a classifier for a link prediction problem. We have used two evaluation metrics to judge the performance of our classifier: 1) Precision recall and 2) Receiver operating characteristic. We calculate the AUC for each metric. The range of AUC is between 0 and 1. AUC value of 1 being the ideal case.

Precision and recall are then defined as:

$$Precision = \frac{tp}{tp + fp} \tag{7.5}$$

$$Recall = \frac{tp}{tp + fn} \tag{7.6}$$

Precision is the probability that the retrieved result is relevant whereas recall is the probability that a relevant result is retrieved in a search.

Results - with Random Removal of Edges

We have extracted three kinds of features from the data: network (local and global) and discussion features. The results for feature ranking on the complete dataset (with missing edges for each case)are stated in Table 7.2. According to the selection methods, Katz coefficient, Cosine similarity and Preferential attachment score are top ranking network features whereas the Smallest discussion size and Largest discussion size are the top ranking discussion features. Overlap coefficient for common discussion is also a good predictor but Overlap coefficient for common neighbors and normalized common neighbors are not adding much information. Therefore, in total we utilize only the top six network features. These results can also be verified from Table 7.3 where we report the AUCROC and AUCPR for each feature using logistic regression and 10-fold cross validation. When there is no network structure available, the network features are not applicable (Table 7.7). We have also reported the performance of two naive models. The first one is a naive random method and the second one assumes a complete mutual friendship graph.

We used logistic regression with 10-fold cross validation on the original imbalanced data for all the experiments, unless stated otherwise. The classification results are plotted in Figures 7.5 and 7.6. Average AUCROC and AUCPR for linear combination of features are stated in Table 7.4. The top features are picked from the results in Table 7.2. The table reports the performance of the classifier as we start adding features. The numbers in first column indicate how many top features for the respective rows are being used for prediction. We used the ranking given by the information gain criteria.

Table 7.3.: Average AUCROC and AUCPR for individual network features using logistic regression and 10-fold cross validation on unbalanced data.

	AUCROC			AUCPR		
Feature name	50%	75%	90%	50%	75%	90%
neighbors	0.802	0.623	0.523	0.527	0.345	0.215
– min-norm.	0.512	0.519	0.509	0.144	0.198	0.191
– max-norm.	0.501	0.504	0.502	0.133	0.174	0.176
Jaccard	0.789	0.620	0.523	0.424	0.313	0.212
Adamic Adar	0.759	0.624	0.523	0.495	0.363	0.218
Preferential attachment	0.743	0.718	**0.629**	0.379	0.391	0.315
Katz	**0.888**	**0.788**	0.613	**0.709**	**0.625**	**0.377**
Cosine	0.733	0.561	0.603	0.420	0.194	0.243

Discussions + network features: When 50% of the edges are missing in the networks, there is a slightly higher AUC for some networks when both network and discussion features are utilized as compared to top-network-only features (Figure 7.5 and Table 7.9). The points are concentrated on the identity line which shows that discussions along with network features are not adding a lot of value. Out of 586 networks about 335 have a higher accuracy. On average both AUCROC and AUCPR have a slightly higher value (Table 7.4). When 75% edges in the networks are missing, 350 networks have higher accuracy when discussion features are added. Discussion features do add some information as demonstrated in the case when 90% of edges in the networks are missing. We get better accuracy with discussion features on more than 400 networks.

The two evaluation mechanisms (ROC and PR) calculate different measures. If the intention is to find out how many 'true positives' are successfully detected by the classifier then ROC is the better choice. However, if one is interested in the precision of the classifier, then PR is the better choice. In PR curves there is a trade off between the precision and recall values. For bigger datasets, as the recall increases, precision may start dropping. Note that if a curve dominates in PR space then it would also dominate in the ROC space (Davis and Goadrich, 2006).

We found that the features are invariant to the underlying classification algorithm. We have only reported the results for discussion features against the naive model using logistic regression, linear discriminant analysis and naive Bayes in Figure 7.6.

Since social networks are sparse, we also tested the performance of the classifier on balanced data. The class imbalance ratio in our original data set is approximately $1 : 10$. We balanced the data by randomly under-sampling the majority class ($0s$). The values of AUC for balanced and

Table 7.4.: Average AUCROC and AUCPR for combination of top (best) features using logistic regression and 10-fold cross validation on unbalanced data.

	AUCROC			AUCPR		
Number of top features used	50%	75%	90%	50%	75%	90%
1	0.888	0.788	0.613	0.709	0.625	0.377
2	0.901	0.783	0.639	0.809	0.622	0.396
3	0.901	0.785	0.639	0.809	0.626	0.396
4	0.914	0.785	*0.686*	0.812	0.626	*0.421*
5	0.918	0.784	*0.699*	0.827	0.626	*0.441*
6	0.918	0.772	*0.699*	0.827	0.605	*0.443*
7	*0.913*	*0.791*	*0.699*	*0.824*	*0.612*	*0.444*
8	*0.916*	*0.791*	*0.701*	*0.826*	*0.616*	*0.448*
9	*0.922*	*0.803*	*0.703*	*0.832*	*0.624*	*0.449*
10	*0.922*	*0.805*	*0.703*	*0.832*	*0.626*	*0.449*
11	*0.922*	*0.806*	*0.703*	*0.833*	*0.627*	*0.451*
12	***0.922***	***0.807***	***0.703***	***0.833***	***0.627***	***0.451***

unbalanced data are plotted against each other in Figure 7.4. For all the 586 networks we found no significant advantage of using balanced data over the original imbalanced data.

Results - with Creating Isolates

The individual performance of network and discussion features is reported in Tables 7.5 and 7.7. The ranking was checked and confirmed using information-gain criteria (Cover and Thomas, 2012).

The selection methods rank Adamic-Adar similarity, number of common neighbors and Jaccard similarity highest among the network features, and smallest and largest discussion size among discussion features.

While Katz similarity actually performs best on an almost complete network, its performance degrades rapidly with the number of excluded neighborhoods. Sarkar et al. Sarkar, Chakrabarti, and Moore, 2011 argue that in social networks longer walks are more relevant if short paths (and 2-hop paths via common neighbors in particular) are rare. The coefficient performs poorly in networks containing 50% or more isolates, whereas local features based on triadic closure still do well. This is in line with previous empirical results on OSNs such as Facebook or Flickr (Backstrom and Leskovec, 2011),(Aiello et al., 2012),(Tarissan, Latapy, and Prieur, 2009).

Table 7.5.: Average AUCROC and AUCPR for individual network features using logistic regression and 10-fold cross validation on unbalanced data.

	AUCROC			AUCPR		
Network Features	50%	75%	90%	50%	75%	90%
neighbors	0.859	0.514	0.501	0.643	0.192	0.165
– min-norm.	0.791	**0.519**	**0.509**	0.544	**0.198**	**0.191**
– max-norm.	0.770	0.504	0.502	0.533	0.174	0.176
Jaccard	0.840	0.514	0.501	0.557	0.188	0.165
Adamic-Adar	**0.865**	0.514	0.501	**0.666**	0.194	0.165
preferential	0.761	0.515	0.503	0.429	0.198	0.174
Katz	0.501	NA	NA	0.174	NA	NA
Cosine	0.780	0.512	0.503	0.378	0.194	0.172

In case no network information is available, no network but only discussion features are applicable (Table 7.7) and we also report on two naïve models for comparison. The first one is a random predictor based on the density of ties in the original network and the second one assumes a complete mutual friendship graph. All discussion features perform better than naïve prediction on average, with higher AUC values on 543 out of 586 networks.

If two friends are repeatedly in discussions which have fewer number of participants, then this is a good indicator of friendship prediction (smallest common discussion size). Larger discussions indicate posts with global interests. The estimate coefficients for the size of smallest common discussion and the size of largest common discussion are negative. This signifies that smaller the size of a discussion, the more exclusive it would be which is indicative of the fact that participants involved in the discussion share friendship ties and they may point that the discussion is of interest to a specific group. Note that we did not look at the topological links between all discussion participants since our focus is the structural features of discussions.

Figure 7.3 shows the distribution of AUCROC and AUCPR for linear combination of features network and discussion features. The accompanying table (7.6) reports the average AUCROC and AUCPR and the deviation from mean. When 50% of the nodes are isolates, there is a slightly higher AUC for some networks when both network and discussion features are utilized as compared to network-only features and discussion features are not adding a lot of value. For more sparse networks, the importance of discussion features gets evident. Discussion features do add significant information as demonstrated in the case when 90% of nodes are isolates and we see more than 20% relative improvement in AUCROC. We noticed that with the addition of a single discussion feature, the AUC improves by 0.101 points. Adding further discussions information results in additional performance improvement. This shows that when the network

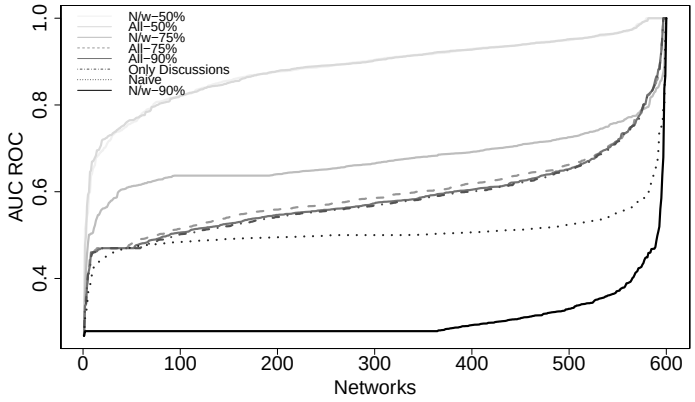


Figure 7.3.: Performance relative to percentage of hidden neighborhoods.

Table 7.6.: Average AUC (and standard deviation) for ROC and PR curves of all network features with and without discussion features.

	AUCROC				AUCPR			
Features	50%	75%	90%	100%	50%	75%	90%	100%
All network features	0.859 (0.09)	0.516 (0.03)	0.503 (0.09)	-	0.647 (0.177)	0.198 (0.130)	0.171 (0.122)	-
All network features + all discussion features	***0.875*** (0.09)	***0.619*** (0.087)	***0.619*** (0.087)	0.606 (0.087)	***0.661*** (0.175)	***0.308*** (0.185)	***0.286*** (0.189)	0.284 (0.189)

is partially available then discussions act as a proxy to detect friendship ties. This strengthens our assumption that even on online social networking sites, people involve more in intra-group interaction.

Discussion Features

Finally we show the case where there is no network information available. In this case we have only discussion features available. We compare the performance of discussion based features and the naive algorithm. Based on the distribution of class labels (0*s* and1*s*) in original data, the naive model predicts class labels. All the discussion based features have a higher value for AUCROC and AUCPR. Out of 586 networks, 543 give higher AUC values with discussions features.

Table 7.7.: Average AUCROC and AUCPR for individual discussion features (without any network structure) using logistic regression and 10-fold cross validation on unbalanced data.

Discussion Features	AUCROC	AUCPR
Joint discussions	**0.608**	**0.284**
– min-normalized	0.606	0.262
– max-normalized	0.607	0.273
Jaccard (discussions)	0.607	0.275
Smallest discussion size	0.598	0.250
Largest discussion size	0.603	0.260
Naïve (random)	0.506	0.173
Naïve (all 1s)	0.500	0.163

7.5. Discussion

Various studies have been conducted on comment mining to predict popularity of stories and news, e.g. on Digg and Slashdot (Potthast et al., 2012),(Jamali and Rangwala, 2009),(Nasim and Brandes, 2014). Studies on using comments information indicate that there is lots information in mining comments on social networking sites. We analyzed the commenting behavior of users to predict links between friends. The results show performance of network and discussion (comments on posts) features in isolation as well as their joint performance. We removed links from the networks to emulate situations where the links are not observable on social network sites. We used a linear combination of the top network features and discussion features. This combination outperforms the performance of the individual features.

In the former sampling scheme, Katz coefficient is the best predictor even when 90% edges from the network are removed. Katz coefficient is the weighted sum of all paths between two nodes. This is a global network feature and outperforms all the local network features. Sarkar *et al.* Sarkar, Chakrabarti, and Moore, 2011 argued that in social networks the number of long paths tend to provide bounds on distance. These bounds are looser than the bounds obtained if enough short paths (e.g. common neighbors) exist. This means that longer paths are more useful if shorter paths are very rare. We also observe the same trend in our results; when the network becomes more and more sparse, the performance of common neighbors decreases but Katz coefficient is still a good predictor in those situations.

While Katz similarity actually performs best on an almost complete network, its performance degrades rapidly with the number of excluded neighborhoods, in the latter sampling scheme,

For more sparse networks, the importance of discussion features slightly gets evident (Table 7.9). For instance when 75% of the edges are missing, the area under ROC curve and area under PR curve see improvement from 0.772 to 0.807 and 0.605 to 0.627 respectively. When an even higher percentage of edges is removed, (as in the 90% case), the area under ROC curve improves from 0.639 to 0.703 and area under PR curve improves from 0.396 to 0.451. Network features together with top discussion features (smallest discussion size and largest discussion size) improve the accuracy by 15%. This shows that structural features of discussions gives similar information as the network structure. With even a little network information available, network features (especially Katz coefficient) are robust in detecting ties. Adding discussions information adds some additional performance improvement. This shows that network features already contain reasonable information. However, when there is no network structure available then we see that discussions act as a proxy to detect friendship ties. This strengthens our assumption that even on online social networking sites, people involve in discussions where there is more chance of intra-group interaction.

If two friends are repeatedly in discussions which have fewer number of participants, then this is a good indicator of friendship prediction (size of Smallest discussion size). Note that we did not look at the topological links between all discussion participants since our focus is the structural features of discussions. Our results also show that smaller the size of the Largest discussion, the more probable a friendship link is. Further, Jaccard index (for discussions) is also a good predictor of friendship ties. It shows the amount of overlap in interests of two people. Another interesting feature which constantly performed at par with the Jaccard index is the Overlap coefficient for discussions. This asymmetric similarity may indicate that one of the nodes is not very active and perhaps comments when the more active node writes on topics of mutual interest.

Because of sparsity of the networks we also experimented with balancing the data. We show

a comparison of performance on balanced and unbalanced data in figure 7.4. It is noteworthy that for this dataset there was no significant difference in average performance of the classifier. Readers are referred to Figure 7.4 where we have plotted AUCROC/AUCPR for balanced data against AUCROC/AUCPR for imbalanced data, for each network. Figures (a) through (d) show AUCROC plots for 50%, 75%, 90% and 100% missing edges respectively and (e) through (h) show AUCPR plots for 50%, 75%, 90% and 100% missing edges respectively . In all graphs, points are concentrated on the identity line which means the results on both balanced and imbalanced data are similar. We finally report the standard deviation for both balanced and imbalanced data in Table 7.10.

The three classifiers: LR, LDA and NB give similar performance. Logistic regression being slightly the better one. LDA assumes that the explanatory variables (features) have a normal distribution. Logistic regression does not have any such assumptions. Cases where the explanatory variables are not normally distributed, LDA does not perform well. We also experimented with support vector machines (SVM). In practice when the data points are very large (typically above $50,000$), the time for computations significantly increases and it is not prudent to use non-linear kernels. For large data, SVM with linear kernel is used Hsu, Chang, and C.-J. Lin, 2003. With linear kernel, SVM algorithm is very similar to logistic regression and performs equally well. We did not see any significant performance difference using SVM on the balanced data(under-sampled).

We posed the question whether it is possible for a third-party application such as Candy Crush or a bot to determine the personal network of a user who is using that application. We find that if as little as 10% of a user's friends have installed the application, it can reveal a significant portion of the user's personal network. We utilized the stylized fact that individuals act as members of multiple social groups. Members of the same group tend to participate in similar activities. Our results are based on multiple network and interaction features as well as multiple classification algorithms, and they suggest that in the absence of network structure, interaction information may be used as a proxy for friendship ties and thereby improve the performance of link prediction.

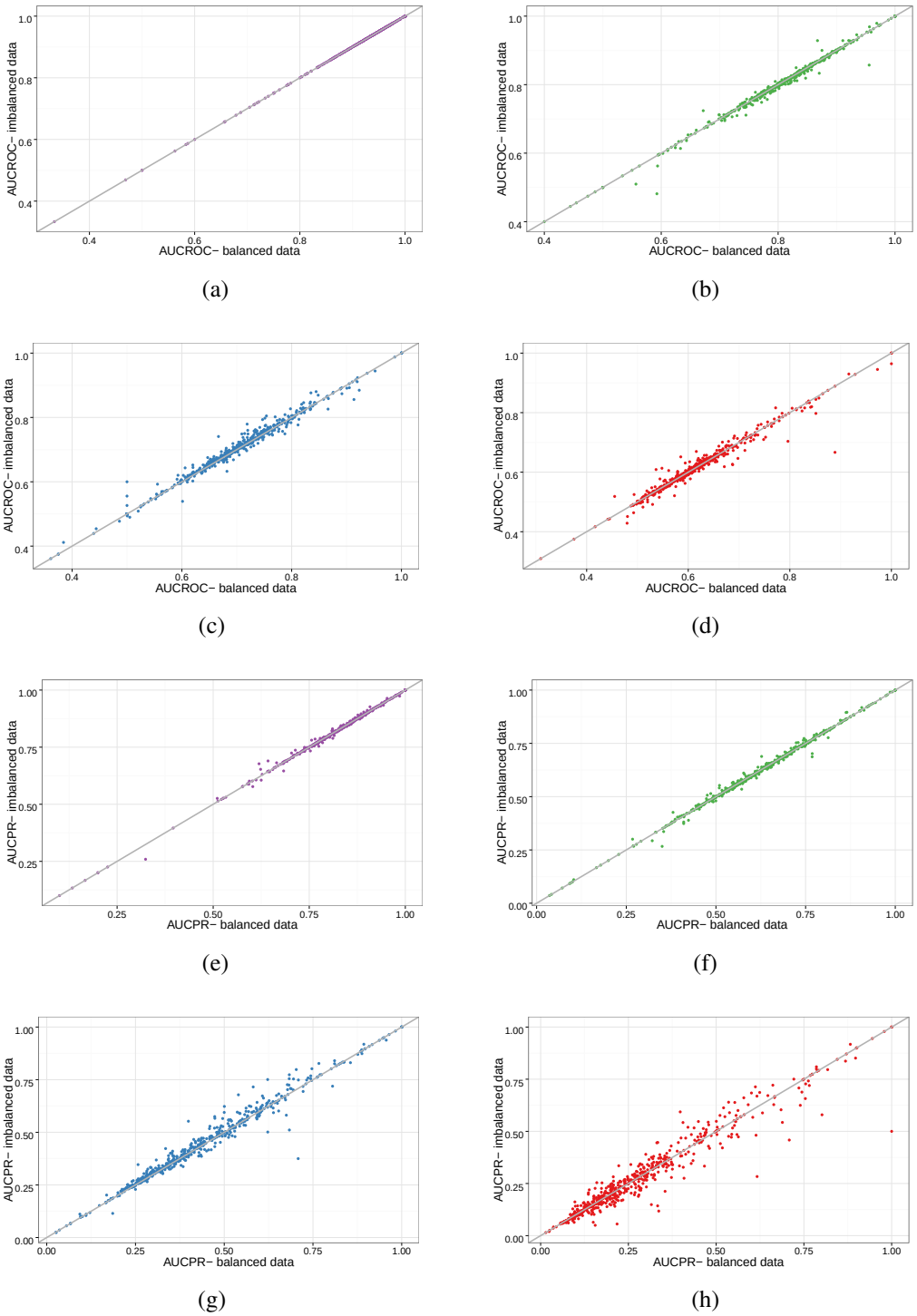
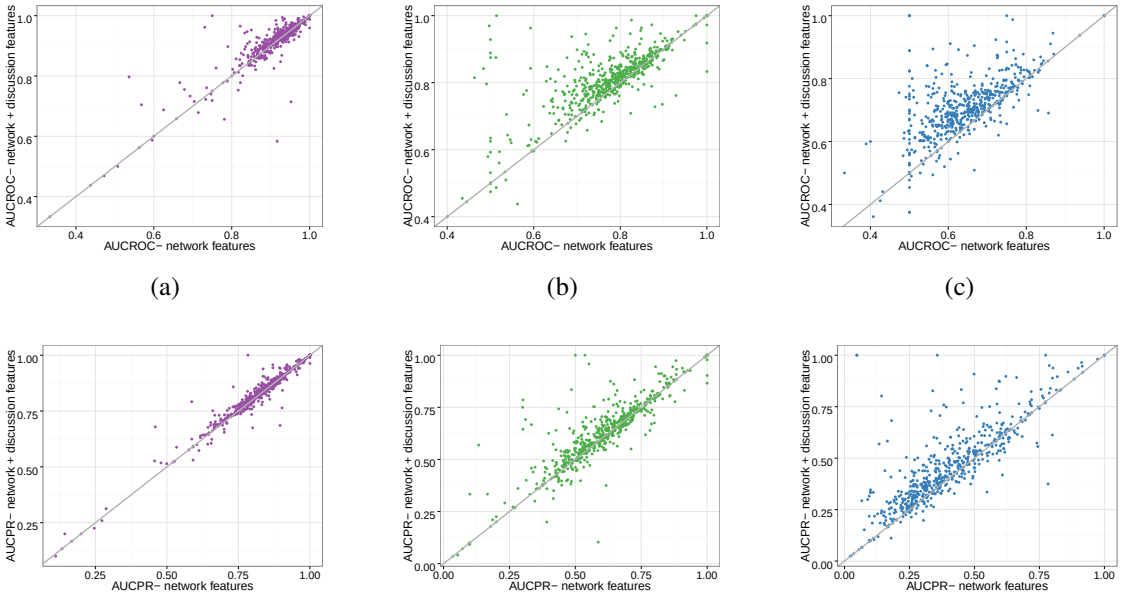


Figure 7.4.: AUCROC and AUCPR for balanced and imbalanced data. Each point represents an ego-network. Graphs are plotted using the discussion features.

Table 7.8.: Standard deviation for balanced and imbalanced data using linear combination of top features (six for the 100% missing edges case).

	AUCROC				AUCPR			
	50%	75%	90%	100%	50%	75%	90%	100%
Balanced	0.071	0.088	0.099	0.089	0.120	0.157	0.188	0.192
Imbalanced	0.074	0.090	0.099	0.091	0.121	0.158	0.190	0.192



(a) (b) (c)

(d) (e) (f)

Figure 7.5.: AUCROC for missing edges: (a) 50%, (b) 75% and (c) 90% . Each point represents results for an ego-network. AUCPR for missing edges: (d) 50%, (e) 75% and (f) 90% .

Table 7.9.: Average AUCROC and AUCPR for top six network features with and without discussion features. Standard deviation is reported in brackets.

	AUCROC			AUCPR		
Features	50%	75%	90%	50%	75%	90%
Top six network features	0.918 (0.076)	0.772 (0.102)	0.639 (0.097)	0.827 (0.123)	0.605 (0.160)	0.396 (0.183)
Top six network features + all six discussion features	***0.922*** (0.074)	***0.807*** (0.090)	***0.703*** (0.099)	***0.833*** (0.121)	***0.627*** (0.159)	***0.451*** (0.190)

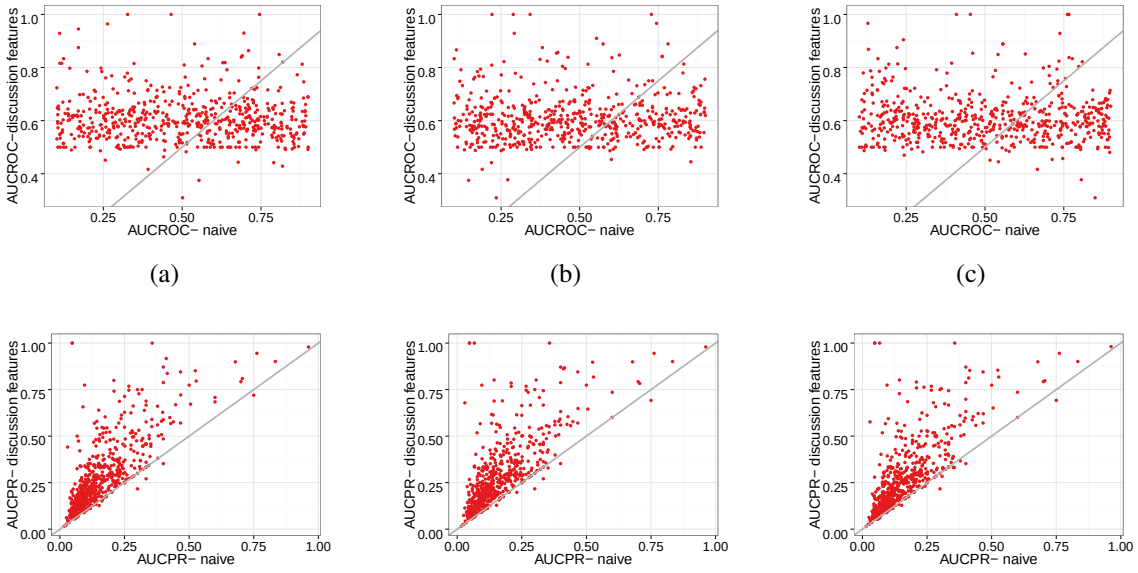


Figure 7.6.: Performance of the classifier when there is no network information available. AUCROC: a) logistic regression b) linear discriminant analysis c) naive bayes, AUCPR: d) logistic regression e) linear discriminant analysis f) naive bayes.

Table 7.10.: Standard deviation using different classifiers with discussion features (100% edges missing).

AUCROC			AUCPR		
LR	LDA	NB	LR	LDA	NB
0.091	0.090	0.089	0.192	0.190	0.192

Part II.

Temporal Regularity in Social Interaction

8. Interaction in Communication Networks

Massive amount of relational event data is generated by social interaction. Such data, as proxy of human relationships is helpful in understanding and predicting behavior of individuals such as influence, activity bursts, buying habits etc. In this part of the thesis we shift the focus to understanding and predicting instantaneous ties in communication networks. In contrast to the first part of the thesis, where we analyzed the interplay between social relations (network ties) and social interactions, in this part we are going to analyze the dynamics of interactions, independent of the network structure. We combine methods from time series analysis and machine learning to explore social interaction patterns.

8.1. Motivation and Background

8.1.1. Periodicity in Human Social Interaction

In this work we focus on mobile calling as an example of demonstrating temporal patterns in communication and also discuss possible benefits of discovering peculiarities in this domain. An estimated 261 million Americans own mobile phones, with a daily average of almost 1.3 billion mobile communication events (Lenhart, 2010). Hence mobile phones represent an important communication medium. A call can be made in a variety of situations because of mobile phone's portable nature and one can assume very little about the context of a call. These two factors, frequency and versatility of use, necessitate an extremely efficient call-making interface design. Users generally make phone calls in two ways: either by selecting the callee from a contact list, or through the call log. The former displays contacts in alphabetical order with no consideration of past calling behavior. While most mobile phones offer the capability of selecting certain contacts as *favorites*, the favorites list is, however, still a static list, requiring active intervention by the user in order to update. Call logs, on the other hand, do take past user behavior into account, displaying called numbers in reverse chronological order. The model of user behavior assumed by call logs is, nonetheless, highly simplistic. It supposes that the

likelihood of calling a particular contact, $P(c)$, is a monotonically decreasing function of the time elapsed since last contact. Sociologists have, however, shown that human life is temporally organized and that most social interactions have fairly reliable temporal regularity (Zerubavel, 1985). This implies that $P(c)$ could be estimated to a certain extent by understanding user calling patterns. Such an implication, correct, would allow for the design of a considerably more efficient calling interface than what is provided by either contact lists, or chronological call logs.

In social network settings, temporal interactions have been used to study human behavior, for instance commenting behavior of Facebook users (a consequence of social selection or social influence effects)(Nasim, Ilyas, et al., 2013). Temporal interactions have also been used to predict links in social networks (Lee et al., 2013; Nasim and Brandes, 2014; Nasim, Charbey, et al., 2016; Tabourier, Libert, and Lambiotte, 2016). Temporal regularity can be observed in time variation of activity on online social networks such as Youtube, Twitter and Slashdot, and also in frequency of edits made on Wikipedia (Gill et al., 2007; Kaltenbrunner et al., 2007; Yasseri, Sumi, and Kertész, 2012). Activity on twitter in various languages shows that circadian patterns exists for tweets all around the world (Thij, Bhulai, and Kampstra, 2014).

Call log data has been shown to hold significant potential of providing insights into the underlying relational dynamics of societies, evolution of relationships over time and, in the absence of survey data, the quantification and prediction of social network structures (Eagle, A. S. Pentland, and Lazer, 2009). MIT Human Dynamics Lab's Reality Mining experiment conducted in 2004 was one of the first experiments to study community dynamics by tracking a sufficient amount of people with their mobile phones(Eagle and A. Pentland, 2006). Data of calling patterns has been used to infer friendships relations and uncover individual and collective human dynamics (Eagle, A. S. Pentland, and Lazer, 2009; Candia et al., 2008; Jiang et al., 2013; Krings et al., 2012). Call-volume data has been used to explore whether the distribution of calls in an urban population follow routine patterns or not, and whether the variation of such patterns in different parts of the city can be explained (Sevtsuk and Ratti, 2010). Several call prediction models have been proposed in the literature. Phithakkitnukoon et al. (2011), predicted the outgoing and incoming calls on Reality Mining dataset of Eagle and A. Pentland (2006), based on most recent calling data. Out of the 94 datasets, they used a small subset of 30 users for performance evaluation. Barzaiq and Loke (2011) modeled the historic call patterns of users and achieved a 35% accuracy for call prediction on synthetic data. Haddad et al. (2014) discuss a probabilistic model that uses call frequency to predict incoming and outgoing calls for each individual contact.Underlying their model is the assumption that the calling behavior of users can indeed be modeled as a periodic phenomenon. The authors tested their model on a large sample by making it available as a mobile application. Recent

studies of human behavior indicate that the timing of communication events is characterized by long dormant periods interspersed with bursts of high activity (Barabasi, 2005; Jo et al., 2012; Y. Wu et al., 2010). Barabasi (2005) attributes this bursty non-Poisson character of human behavior to a priority-based queuing process. This view is supported by Jo et al. (2012) who show that burstiness remains in mobile communication data even after circadian and weekly patterns have been removed, precluding the attribution of periods of inactivity to nights or weekends. They conclude that burstiness results from non-homogeneity in human task execution mechanisms. H. Kim, Zang, and Ma (2013), conducted a study on a large dataset from North-American mobile phone users. The results suggest that the caller-callee behavior cannot solely be modeled using the Poisson distribution. Based on frequency of information exchange between the users, they classified the user-pairs into three categories characterized by the inter-arrival times between calls made between pairs. Obtaining information about the family/friends is difficult for a scenario where the aim is to predict the next call for users belonging to the general population.Inspired by effective studies on calling patterns, researchers have devised several call prediction models.

In a related study, Cardillo et al. (2014) studied human proximity patterns in two data sets: the Reality Mining dataset and the co-location traces from INFOCOM'06. They found that proximity patterns from the MIT data contain both weekly and daily periodicity - most probably a result of how academic activities are scheduled at a university - while the INFOCOM'06 data showed only daily periodicity. Caridillo et al. extended this observation to study how cooperation emerges in a human society.

While studying social network turnover,T. Aledavood, E. López, et al. (2015b) and T. Aledavood, E. López, et al. (2015a) found that individual calling and messaging behavior follows a circadian rhythm. Their study of 24 subjects revealed that the frequency and entropy of communication displays a distinct daily pattern that remains persistent over time. Moreover, it was found that frequently called contacts are the ones most likely to be contacted during low entropy periods. Nonetheless, the study does not answer the question whether communication between pairs of individuals is periodic. Further, findings on temporal patterns in (Talayeh Aledavood, Lehmann, and Saramäki, 2015) are attributed to the diurnal cycle of human beings.

A patent from Google suggests that an adaptive contact list may detect contextual information for a given mobile phone user and may identify appropriate contact entries Google Inc (2010). While studying the effects of two different UI adaptation techniques on user performance, Tsandilas (2005) conclude that adaptation is always more effective, even when the accuracy of prediction is low. Bentley and Chen (2015) found that the majority of contacts in a modern aggregated mobile phonebook are rarely used. Their study shows that the five most frequently contacted alters represent 80% of phone and text communication. In addition, they found that

a median of $Q = 60\%$ (six out of a total of $N = 10$ contacts) displayed in a "recent calls" list are amongst the most frequently contacted. While the authors use this latter statistic to argue against the efficacy of a "recent calls" list, it would be interesting to explore whether the value of Q increases for larger values of N, especially since the authors' results indicate an upward trend in Q as N increases from one to 10. Proponents of a "recent calls" list may argue that, in practice, these lists hold more than 10 entries. Based on their results, Bentley and Chen suggest a redesign of the content and representation of contact lists. A redesign of contacts book was proposed also in (Oulasvirta, Raento, and Tiitta, 2005). The data for Bentley and Chen's study was collected from user in the United States via an Android app. Volunteer bias especially as a survey was also required from the users. Moreover, while representative of the general population of the US, the authors acknowledge that communication patterns in other parts of the world may vary.

In this work we provide the data analytics and then discuss findings related to the daily and hourly patterns in the communication of smartphone users. We start at an aggregate level, studying aggregated patterns and discussing our findings. We then move on to the individual level, and focus on the variation that is hidden at the aggregate level, i.e. individual differences and marked daily activity patterns.

Contribution: Our contribution in this chapter is as follows:

1. Collection of an original mobile phone usage dataset of 783 users with 229,450 communication events. The dataset captures call data from an understudied population: Pakistani mobile phone users [1].
2. Analysis of the collected data to answer the research questions posed in the last section. We study calling patterns of individuals both at aggregate level and at a finer level, i.e. pairs of ego and alters.

8.2. Data Collection

We began this work by conducting a small scale pretest. We first collected online survey-responses from 28 participants who own Android phones and then analyzed call data from 13 of those participants (Data statistics are available in Appendix A). The survey aimed at getting a deeper insight into the calling behavior of users with respect to the use of call logs for making future calls. Survey consisted of questions such as: whether the participants call different people

[1] The responsibility of data collection, anonymization and storage lies with COMSATS institute of Information Technology, Pakistan, under letter number: CIIT/CS-SP-15/ISB. My involvement was limited to the design phase of the application.

on different days of the week; whether he/she calls different people on weekends; whether they use missed calls to indicate any kind of signal; how often they use call log to dial a number, etc. Once the participants finished the survey, they were given the option to share their data for research purpose. Only 13 participants agreed to share the data. We then sent a link to our app hosted on Google Play to those participants. Once the app was installed, it automatically submitted an anonymized version of the data to our server. Most survey participants agreed that they usually use call logs to make future calls. Except for four participants, everyone else had experience in using missed call as a signal, for instance as an indication for the other person to callback. More than 50% of the participants agreed that they call different people on weekends as opposed to weekdays. Analysis of their call log data indicated that 30% ego-alter [2] pairs communicated more on weekends as compared to the weekdays.

Encouraged by the qualitative as well as the quantitative results of the pretest, we collected a dataset from the general population to get further insight into the calling behavior. We were interested in a number of research questions related to understanding the characteristics of communication behavior on mobile phones, and patterns of communication at individual level as well as at a finer level i.e. ego-alter interaction. We were interested in knowing: how many active contacts do mobile phone users have? how often they are called? With respect to historic logs, we were interested in finding: Distribution of calls, more specifically, what percentage of communication goes to top contacts? and how often people call the recently called contacts?

We developed a smartphone app specifically for the purpose of this experiment. Data collection was limited demographically to users of smartphones running the Android OS, and geographically to the country of Pakistan. The data was collected by marketing the application through Facebook ads, word of mouth and wall posts on technology pages on Facebook. Pakistan ranks 8^{th} in the list of number of mobile phones in use in the country, ahead of Japan, South Korea and all European countries(Wikipedia, 2016). In December 2015, the number of 3G/4G subscribers exceeded 23.16 Million (PTA, 2016). While industry sources estimate that Android users represent 68% of the total smartphone population in that country, extensive market surveys are lacking and, hence, conclusive judgments about the qualitative nature of the sample cannot be made.

Pakistan is a low income country and people are interested in reducing their mobile usage cost (Agüero and De Silva, 2009). To make its value proposition more attractive, the application presented users with the most economical telecom service for their needs based on past calling behavior. These telecom services - also referred to as “packages” in the local parlance - differ primarily in the calling rates they offer during specific hours of the week. A recommender

[2]Ego is the focal actor who has installed the app. Alters are people with whom ego communicates using voice calls.

system for similar telecom products was developed by Zhang et al., 2013. But, where they used fuzzy-set techniques to select the most economical product, our recommendations are based on a simple simulation run with the users' call history. Including this additional functionality in our data-gathering app not only expanded our sample set, but we also expected it to mitigate the volunteer bias natural in such survey data collection methods. Users were notified that their call data would be used for academic research purposes.

Since the data was collected on mass scale, collecting demographic information at that scale was not feasible. We call this data as the *smartphone dataset*.

8.3. Aggregated Data Analysis

We uploaded our application on Google Play on July 28, 2015. The process of data collection lasted from July 28, 2015 till September 24, 2015. The application did not replace any functionality on the host phone and did not interfere with normal usage of phone in any way. The application collected the historic data from call logs of smartphone users. Data consisted of the following fields (anonymized): unique id of the phone, contact that was communicated with, communication event type, i.e. received call, missed call, outgoing call, etc., date and time of the event and duration of call. The data collected by the application had timestamps from April 19, 2015 till September 23, 2015. The data consists of 783 users with 229,450 communication events and more than 12,000 active contacts. About 83% of the communication took place between 6:00 a.m. and 9:00 p.m.

Of the total calls, 24% calls were incoming and answered, 19% were incoming and missed calls, and 54% of calls were outgoing calls. These call statistics are very interesting, firstly, because Bentley and Chen Bentley and Chen, 2015 also reported similar statistics for their dataset of 200 users, and secondly, they are comparable with the statistics of the 13 users from our pretest. We also looked at the statistics from the Reality Mining dataset that was collected at MIT. There were 22% incoming calls and 66% outgoing calls. However, the percentage of missed calls was only 10%. We also found that 2% calls were incoming and rejected calls in our dataset. None of the calls were voice mail calls, whereas, a minuscule number of calls were from the refused list. The average length of incoming phone calls was around 104 seconds with a median value of 38 seconds. Average length of outgoing calls was 56 seconds. A handful of calls were very long. The longest call lasted one hour. Only 1.6% calls lasted more than 10 minutes. About 83% of the communication took place between 6:00 a.m. and 9:00 p.m. It was especially interesting to note the relatively high percentage of missed calls. Although, a high number of missed calls was also reported in Bentley and Chen, 2015 but the authors could not mention a plausible reason behind it. This statistic for missed calls is a noticeable

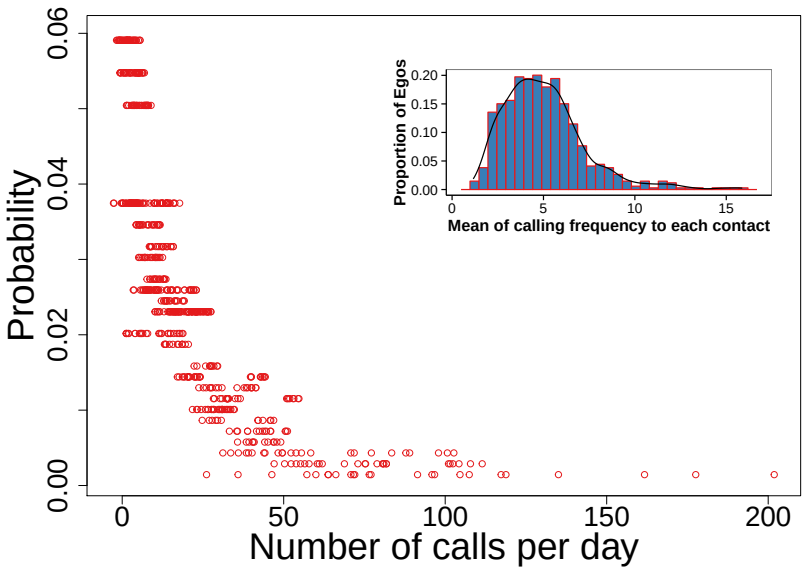


Figure 8.1.: Average number of calls per day and the overlay plot at the top right shows mean of average number of calls per user per contact.

artifact of our dataset from a low-income country, where users, sometimes give missed calls to indicate a signal - an easy way to save money. One of the participants in our user study stated, "*Whenever I reach home late, I give a missed call to my mother so that she could open the door*". Another participant indicated that he regularly talks to his girlfriend in the evening. *"When her parents are around, she gives me a missed call which means that I am not to supposed to call her"*(a very typical setting in a South-Asian society). The notion of missed calls is in the core of Pakistani mobile phone market to such an extent that Unilever launched a campaign called "missed-calls" on mother's day. The campaign attracted over 1 Million calls in 2015 (mmaglobal, 2015).

8.3.1. Distribution of Calls

Bentley and Chen (2015) observed that most calls are to 5-10 of contacts. Similarly, Bergman et al. (2012) observed that the participants of their study did not call 47% of their contacts for 6 months. There is a general observation that there are fewer contacts who are called more often and a lot many contacts who are less often. We observed that the outgoing calls are not normally distributed. Mobile phone users have various kinds of contacts in the contacts list such as friends, acquaintances, family members, workplace contacts, services related contacts, etc., (Bergman et al., 2012). Some of these contacts are frequently called, others are occasionally

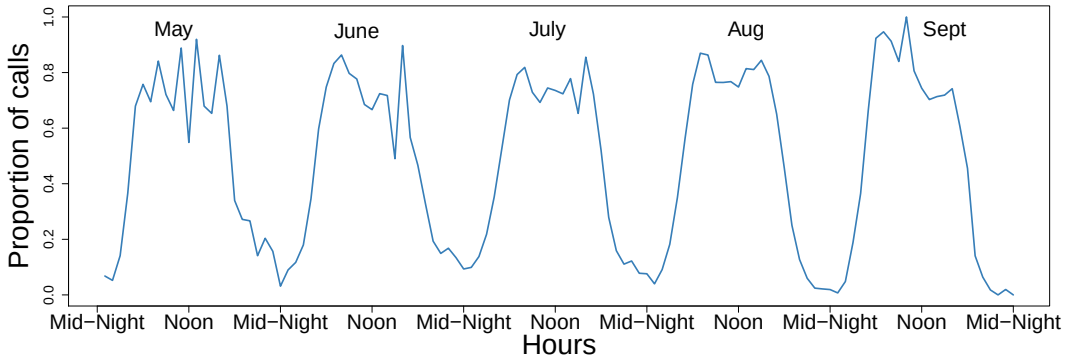


Figure 8.2.: Number of events per hour for each month in the Dataset at an aggregated level.

called and some contacts are not called in a long time. This view is also supported by H. Kim, Zang, and Ma (2013).We empirically found a similar but more interesting pattern; every user's call distribution very closely follow the equation below:

$$\frac{e^a}{x^b} \tag{8.1}$$

Here, a and b are real number that is fixed for each participant and x is the rank of the alter that varies from 1 for the alter with the most communication events and so on till the rank of the alter with the least communication events. It is worth noting that a and b both lie in a narrow range as a varied between 0-7 and b varied between 0-2.5. We observed that our equation fits the the data very well and we got a mean adjusted R^2 of 0.89 and the standard deviation was 0.16. Note we removed the data of 27 egos because their number of communication events was below 4. It is interesting to note that cities and their rank also follow a similar distribution and this pattern is generally known as the *rank-size rule* Rosen and Resnick, 1980. Equation 8.1 indicates that any future redesign of the contact list would probably need to compute the important contacts for each individual. The number of important(top) contacts can vary from about 5 for an individual with $a = 2$ to about 20 for an individual whose value of a is 6. We plan to further investigate why the Equation 8.1 varies from one individual to another and then apply that knowledge to improve calling experience of mobile users. The probability distribution function (PDF) of number of calls per user are shown in Figure 8.1. On average, each user made or received ≈ 22 calls per day. The mean of average number of calls per user per contact is plotted as a bar chart at top right in the same Figure.

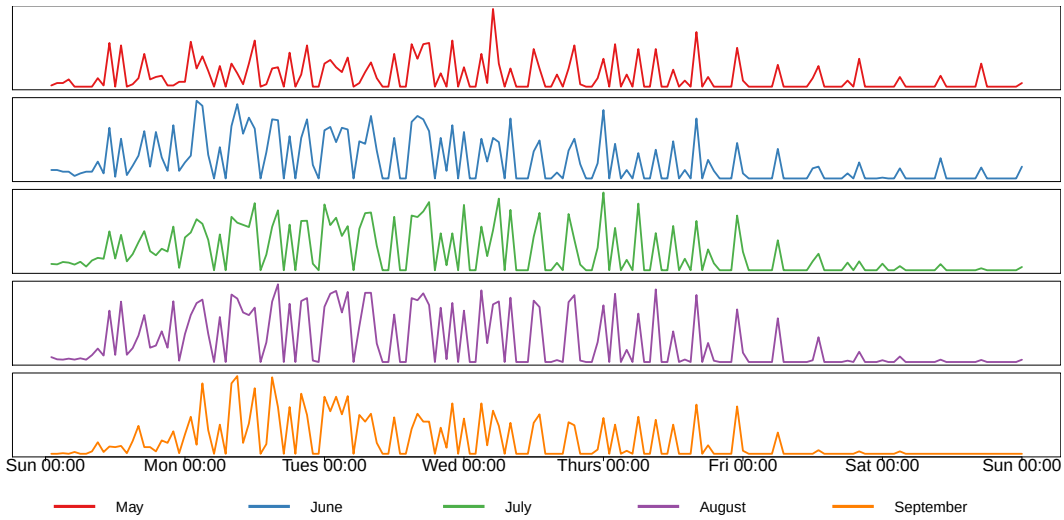


Figure 8.3.: This Figure shows the calling patterns between Sunday and Saturday for each Month. The x-axis depicts the number of hours in each day for 7 seven days of the week. The difference between two ticks on x-axis is 24 hours (1 day). y-axis represents the aggregated proportion of calls made at each hour of the day by the egos. The distributions across all months look identical. Highest communication activity is observed between Monday and Thursday. Call volume decreases significantly, after Friday midday. Comparing all the distributions $\binom{n}{2}$ using Wilcoxon signed-rank test gave the following *p-values*: $0.96, 0.70, 0, 89, 0.41, 0.68, 0.85, 0.34, 0.78, 0, 20, 0.27$.

8.3.2. Hourly and Weekly Calling Behavior

Our data consists of communication events between April and September 2015. Since, there were only eight events recorded in the month of April, we illustrated the aggregate number of events per hour for each month, from May till September in Figure 8.2. The communication activity is highest during the daytime hours and decreases by mid-night in every month.

We also conducted the *Wilcoxon signed-rank test*, which is a non-parametric statistical hypothesis test where our null hypothesis was that the distribution of weekly calls for any pair of months is identical (Figure 8.3). The results show that the *p-values* are not statistically significant, hence insufficient evidence to reject the null hypothesis at $\alpha = 0.05$. This test is useful when the population cannot be assumed to be normally distributed. Only 26% calls were made on weekends in the entire dataset. A low activity during the weekend was consistent for all months.

At a finer level we found that there was a higher probability of communication between 25% ego-alter pairs on weekends. Similarly, 75% ego-alter pairs were more likely to communicate during the weekdays in the dataset. This indicated that probably the nature of the ego-alter social relation is different for those people that a person calls on weekends versus those which he/she calls on weekdays. In the next section we further investigate the dyadic interaction

patterns.

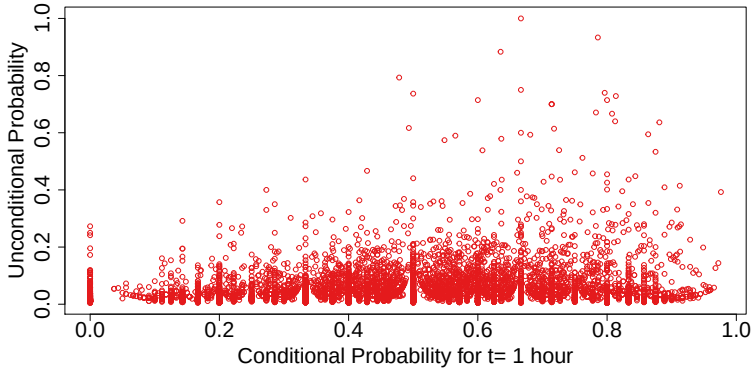
8.4. Ego-Alter Interaction Patterns

Latest approaches on network analysis are exploring the conceptualization of networks as collection of data on dyads rather than on the graph. *"The distinctive characteristic of networks is that their units of observation (the identifiers of data points) are not single entities but pairs of entities, and that each entity may appear in multiple such pairs"*,(Brandes, Robins, et al., 2013). Studies on human communication behavior (Talayeh Aledavood, Lehmann, and Saramäki, 2015; Gill et al., 2007; Thij, Bhulai, and Kampstra, 2014; Yasseri, Sumi, and Kertész, 2012), have investigated circadian patterns. They attributed the temporal regularity to diurnal cycle of human beings (when people are using modern communication modes such as mobiles and Internet) and that inter-individual differences arise due to geographical and cultural differences. We argue that the temporal communications patterns are not just a consequence of the diurnal cycle. At a finer level, we focus on the variation that is covert at the aggregate level, i.e. individual differences and marked daily activity patterns between pairs of users.

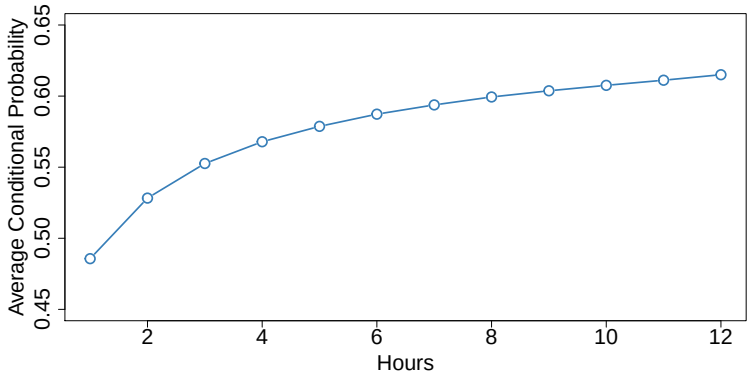
8.4.1. Probability of Calling an Alter Again

Our informal user study showed that 71% of respondents either always or usually use call log to initiate a call. If this trend is true in general, then the probability of a communication event between an ego-alter pair should be significantly less than the conditional probability of a communication event between an ego-alter pair given that there was a communication event in the near past. Since we could not come up with a reasonable definition of near past, so we decided to check this hypothesis by computing the conditional probability between each ego-alter pair given that there was a communication event t hours ago, where $t \in \{1, 2, \cdots, 12\}$. Figure 8.4a clearly showed that a communication event is much more likely if there was a communication event within the last 60 minutes. However, this conditional probability does not increase significantly and levels off very quickly as we increase t as shown in Figure 8.4b. Moreover, Figure 8.4c shows a histograms of unconditional as well as conditional probability when $t = 1hour$) for all ego-alter pairs. The distribution for conditional probability is normal and hence we estimate the mean and standard deviation ($L(\mu, \sigma^2)$) of this distribution using the Maximum Likelihood Estimate. We estimated $\mu = 0.48$ and $\sigma^2 = 0.21$ as compared to the unconditional probability where $\mu = 0.041$ and $\sigma^2 = 0.059$. To the best of our knowledge, we interpret these results as the first empirical evidence that call logs are fairly useful for making future calls. These descriptive statistics are also consistent with the survey results that we

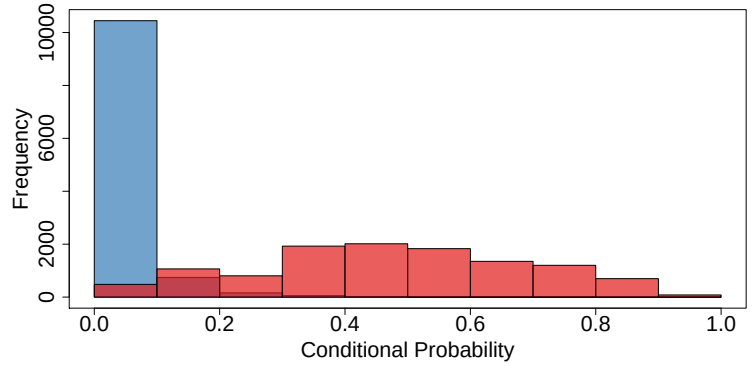
discussed in the Introduction.



(a)Probability of a communication event between an ego-alter pair when there was another communication event between the same pair. For almost all pairs it is clearly greater than unconditional probability of a communication event between that pair.



(b)The predictive power of conditional probability increases very gradually and levels off very quickly.



(c)Distribution of unconditional probability (blue) and conditional probability (red) for all ego-alter pairs for $t = 1hour$

Figure 8.4.: Weekly and Hourly Calling Activity in each Month.

8.4.2. Autocorrelation

Time domain periodicity detection methods make use of autocorrelation functions (Parthasarathy, Mehta, and Srinivasan, 2006). Autocorrelation refers to the statistical dependency between the values of a variable on related entities. In terms of time series data, like our dataset, autocorrelation implies persistence from one observation to another. Autocorrelation is a common characteristic of relational and social-network datasets; since mobile calling is also a form of social interaction, it is plausible to test whether caller-callee interactions exhibit autocorrelation or not. In many time series, it is plausible to expect that the m recent data points are likely to have an influence on the future data points

For comparison purpose, the sample for this experiment consisted of two datasets: the Smartphone dataset and the Reality Mining dataset of Eagle and Pentland (Eagle and A. Pentland, 2006) collected at the Massachusetts Institute of Technology. This latter dataset comprises call and text data for 94 egos. Each dataset was analysed independently. The data for each ego was grouped according to the contact the communication event was initiated to. We represented the individual communication events between the pair using a binary string, with each bit position representing a time quantum, modelled by a Bernoulli random variable Z. A bit (Z) was assigned a value of one if a communication event did occur during that time quantum, and zero otherwise. We have taken into account two types of time quantum —daily and hourly. Studies on human social behavior support our selection of time quantum. Human circadian rhythms intrinsically follow a period of approximately 24 hours (Wever, 2013). Within this broad period, however, there are significant inter-individual differences which may correlate with distinctive temporal patterns of physiological and psychological variables, of gender and, personality traits (Tsaousis, 2010).

We then determined temporal regularity for ego-alter pair using the Ljung-Box Q test ($\alpha = 0.05$) on the string. The Ljung–Box test, also known as a *portmanteau* test, is a function of the accumulated sample autocorrelations r_k, up to any specified time lag m (Ljung and Box, 1978). The autocorrelation was checked up to six lags. If the lag is too small, the Q-test may not detect serial correlation at high-orders. If its too large, the test may have low power since the significant correlation at one lag may be diluted by insignificant correlations at other lags. Computations were performed using the PerformanceAnalytics library in RPeterson and Carl, 2014. As a function of m, it is determined as:

$$Q(m) = n(n+2)\sum_{k=1}^{h}\frac{r_k^2}{n-k} \tag{8.2}$$

where n is the number of data points.

The null hypothesis for Ljung–Box test states that the data is independently distributed (any

	r_{daily}	r_{hourly}	$r_{7am-8pm}$
Reality Mining	0.55	0.89	0.65
Smartphone	0.15	0.60	0.44

Table 8.1.: Proportion of ego-alter pairs demonstrating significant autocorrelation at different time periods

observed correlations in the data is a result of the randomness of the sampling process). The alternate hypothesis is that the data is not independently distributed; it exhibits autocorrelation. The p-value is used to decide if data points are not independently distributed. Typically, when the p-value is less than 0.05, the null hypothesis for Ljung–Box test is rejected.

For each ego-alter pair, an hourly and a daily autocorrelation measure was calculated using the Ljung–Box test where a $p - value < 0.05$ means there is autocorrelation. Table 8.1. lists the proportion of ego-alter pairs that displayed autocorrelation in each of the two datasets. As the hourly autocorrelation measure may have been biased by lack of activity at night hours, a third autocorrelation for communication events between 7am and 8pm was also calculated. For each ego, we removed the alters who were communicated less than the mean of the communication frequency for that ego. This was plausible since on average, the mean lied in the upper quantile (with $\sigma > \mu$) of the calling distribution. The number of top contacts ranged between 5 and 20. This way we also removed the sparse data having insufficient number of communication events required to determine the autocorrelation.

As compared to our dataset, a higher percentage of ego-alter pairs in the reality mining dataset exhibit a daily as well as hourly periodic calling behaviour. The Reality Mining dataset was collected almost a decade ago when other means of smartphone communication such as Whatsapp, Viber, Facetime, etc. did not exist. In the smartphone dataset, we find that a small percentage of ego-alter pairs exhibit daily temporal autocorrelation. This might be an artifact arising from the shift in communication from mobile phone calls/text messages to smartphone instant messengers. Another tenable explanation could be the bias in the datasets. Contrary to the Reality Mining dataset that contains data from students or faculty of MIT media lab with daily activities structured around the academic calendar, the smartphone dataset contains data from general population of a developing country. Further, this is also an indication that communication patterns in different parts of the world may vary which is also acknowledged in (Bentley and Chen, 2015) and thus it justifies the need to study the understudied populations that have a significant mobile phones user base. Notwithstanding a low proportion of time series exhibit autocorrelation in the daily interaction of Smartphone data, there is indeed an indication of periodic calling at finer levels.

8.5. Discussion

Mobile phones represent one of the most commonly used communication medium. The portable nature of the medium means very little can be assumed about the situation in which the phone is used; a typical user makes calls in all kinds of contexts. These two factors, frequency and versatility of use, necessitate an extremely efficient call-making interface design.

As a first step we conducted a pretest on two mobile phone datasets for determining whether users have a regular calling pattern or not. We modeled the communication between mobile phone users as a time series data analysis problem.

We collected and analyzed call data of Pakistani users which is an understudied population. Prior to collecting data on mass scale, we conducted a pretest by first collecting survey-responses from 28 participants and then analyzing the data from 13 of those participants. Our pretest indicated that people often use call logs for making future calls. Encouraged by our initial findings, we launched an android app and collected data from general population for a large scale analysis. On an aggregate level our data statistics were surprisingly comparable with the ones from Reality Mining Eagle and A. Pentland, 2006 and Bentley and Chen Bentley and Chen, 2015. We analyzed daily and weekly temporal patterns, showed that distribution of calls in an ego profile follows the rank size rule, detected periodicity at ego-alter pair level using autocorrelation and compared the results with the Reality Mining dataset. Further, we empirically observed that call logs are an efficient way of dialing future calls. We also deliberated on the rationale behind high percentage of missed calls in our dataset.

In many time series, it is plausible to expect that the recent data points are likely to have an influence on the future data points. In order to identify whether the ego-alters communication data has a pattern, we used autocorrelation which is a type of correlation statistic specifically for correlating the recent data point to other data points in the series. The results on two different datasets quantitatively show that a reasonable number of ego-alter pairs exhibit autocorrelation.

The results show that more than 50% of ego-alter pairs in the reality mining dataset exhibit a daily as well as hourly periodic calling behaviour. The Reality Mining dataset was collected almost a decade ago when other means of smartphone communication such as Whatsapp, Viber, Facetime, etc. did not exist. In the smartphone dataset, we fail to find daily temporal autocorrelation. This might be an artifact arising from the shift in communication from mobile phone calls/text messages to smartphone instant messengers. Another tenable explanation could be the bias in the datasets. Contrary to the Reality Mining dataset that contains data from students or faculty of MIT media lab with daily activities structured around the academic calendar, the smartphone dataset contains data from general population of a developing country. Notwithstanding that a low proportion of time series exhibit autocorrelation in the daily interaction of Smartphone data, there is indeed an indication of periodic calling at finer levels.

The importance of intuitively sorting communication events (on displays) entails constant improvement in the user interface of interactive products, services, or systems. For instance, 'Google Inbox', a recent service from Google, is an example of an instinctual interface that accurately sorts communication events, in this case the emails through an intuitive interface that makes it very easy to take action on important emails. As mentioned in the start of the paper, smartphones often store a log of communication(e.g., phone calls) in what is commonly called a call log. Information in the call log may selectively be displayed to the user of the smartphone. Such information is useful to the user, among other things, to make future calls. When the users follow a temporal communication pattern, the sometimes tedious interface through which calls are initiated can be improved. In an independent study Bentley and Chen also reported that while 'recents-list' could be good for contacting "temporary" contacts which are perhaps relevant over the course of a day, this list does not provide easy access to the most frequently contacted people. Our results imply that in most cases ego-alter interactions do have temporal patterns. Patterns in time series can be used for forecasting the future calling behavior of users. Identifying idiosyncrasies in the ego-alter communication can help improve the calling experience of smartphone users by automatically (smartly) sorting the call log without any manual intervention. We believe this study would serve as a precursor for conducting large as well as controlled studies, to find the subtleties behind ego-alter communication, with the purpose of improving the calling interface by making the call logs more efficient and user friendly.

9. Call Prediction Using Temporal Information

9.1. Time series analysis

A *time series* can be thought of as a sequence of observations over time. It can be defined as follows:

Definition A *time series* T is an ordered sequence of n-real valued observations.

$$T = (t_1, \ldots, t_n), t_i \in \mathbb{R}$$

Most often time series is a result of the observation of an underlying process where values are collected over time at uniformly spaced time instants, according to a given sampling rate (Esling and Agon, 2012).

Time series can cover the full set of data or may contain a subsequence. It is defined as follows (Esling and Agon, 2012):

Definition Given a time series $T = (t+1, \ldots, t_n)$ of length n, a *subsequence* S of time series T is a series of length $m \leq n$ consisting of contiguous time instants from T.

$$S = (t_k, t_{k+1}, \ldots, t_{k+m-1})$$

where, $1 \leq k \leq n-m+1$.

Time series analysis can be used for prediction of future events based on the past, controlling the process producing the event, understanding the mechanism that is generating the time series and identifying the salient features of the series. Predicting a time series is possible if and only if the dependence between values existing in the past is preserved also in the future (Bontempi, n.d.). Time series analysis has applications in Meteorology, Finance and Economics, Marketing, and Medicine,to name a few.

In the data mining domain, clustering has been used to detect patterns in time series. The task of clustering a time series can be of two types: *whole series clustering* and *subsequence clustering*.

Whole series clustering can be defined as follows:

Definition Given a time series database *DB* and a similarity measure $\mathcal{D}(\mathbf{Q},T)$, find a set of clusters $C = \{c_i\}$ where $c_i = \{T_k | T_k \in DB\}$, which maximized that inter-cluster distance and minimizes the intra-cluster variance.

where *DB* is an unordered set of time series and the similarity measure $\mathcal{D}(\mathbf{Q},T)$ between time series Q and T is a function taking two time series as input and returning *distance* between them.

Subsequence clustering is defined as follows:

Definition Given a time series $T = (t_1,\ldots,t_n)$ and a similarity measure $\mathcal{D}(\mathbf{Q},C)$ find a set of clusters $C = \{c_i\}$ where $c_i = \{T_j | T_j \in \mathbf{S}_T^n\}$ is a set of subsequences that maximizes inter cluster distance and intra cluster cohesion.

9.2. Communication Networks

An estimated 261 million Americans own mobile phones. A typical American makes an average of 5 calls in a single day (Lenhart, 2010). With a daily average of almost 1.3 billion communication events and an annual total of 2.45 trillion minutes of usage (CTIA, 2015) in the US alone, mobile phones represent one of the most commonly used communication medium. Judging from the trend in recent years, the user base of cell phones can be expected to further increase in the future. Hence, a deeper understanding of the temporal structure of mobile phone communication would allow us to optimize and streamline a technology that has penetrated the very fabric of human life.

Systems with time-stamped dyadic interactions can be modeled as temporal networks. Time stamped networks of human communication along with proximity networks, are the largest class of systems modeled as temporal networks. Sociologists have shown that human life is temporally organized and that most social interactions have fairly reliable temporal regularity (Zerubavel, 1985).

When the dependence between interaction values in the past is preserved in the future, then future interactions are reasonably easy to predict. In numerous studies((Jo et al., 2012), (H. Kim, Zang, and Ma, 2013)) inhomogeneity has also been observed in human activities. T. Aledavood, E. López, et al. (2015a) showed that the individual differences in the distribution of calling remain persistent. They also suggested that frequently called contacts are the ones most likely to be contacted during low entropy periods. Furthermore, causal events are also a key characteristic of human communication behavior. Nasim, Rextin, Khan, et al. (2016) showed that communication behavior varies between pairs of users. This unique mix of characteristics make

the prediction of future interactions all the way more challenging. Typically, communication networks data comes from mobile phone call logs or email networks.

Mobile phones can collect traces on human activity patterns including information about user's location,interaction timings, interaction patterns etc. Call log datasets have been explored in the past. Bentley and Chen (2015)and Nasim, Rextin, Khan, et al. (2016) observed that datasets from different geographic regions may show distinctive communication traits. In this work, we reuse the data collected by Nasim, Rextin, Khan, et al. (2016) (we call it the Smartphone dataset)from an economically underdeveloped country. We test the performance of our calls prediction classifier on the dataset of Eagle and A. Pentland (2006), in addition to the Smartphone dataset. Based on the assumption that one could predict the users' calling behavior using temporal features, we first understand the mechanism that generates a series of events between an ego-alter pair; we then identify the salient features of the series and finally using hybrid machine learning approach predict the future communication events.

Contribution: Our contribution in this chapter is as follows:

1. In this work we explore temporal homogeneity/non-homogeneity in mobile phone calls, in order to predict future communication events between pairs of individuals. We perform a study of possible features in time series analysis that are useful in call prediction. Using actual call logs we show that majority of users are not optimally served by existing calling applications such as call logs. Further, we also test the hypothesis that the majority of caller-callee interactions display temporal regularity through a statistical measure called autocorrelation. We then model the call prediction problem as a supervised multiclass classification problem and test the usability of these features using a machine learning approach. In particular we explore calling frequency, recency of calls, and temporal regularities for improving the prediction of calls.
2. Lastly we perform experiments on both the collected data as well as on the famous Reality Mining Dataset (Eagle and A. Pentland, 2006) to demonstrate applicability of our methods for predicting future calls.

Problem Statement

Let U be the set of all egos in the dataset. For every $u_i \in U$, let X_i be the set of contacts. Formally, we define the call prediction problem as follows: Given the historical communication events, $\{Y_i(t_0),...,Y_i(t_{(n-1)})\}$ consisting of outgoing, incoming and missed calls that occurred at time $t_0,...,t_{n-1}$, for a user $u_i \in U$, predict which contact$\{x_1,...,x_m\} \in X_i$, u_i is going to call at time t_n.

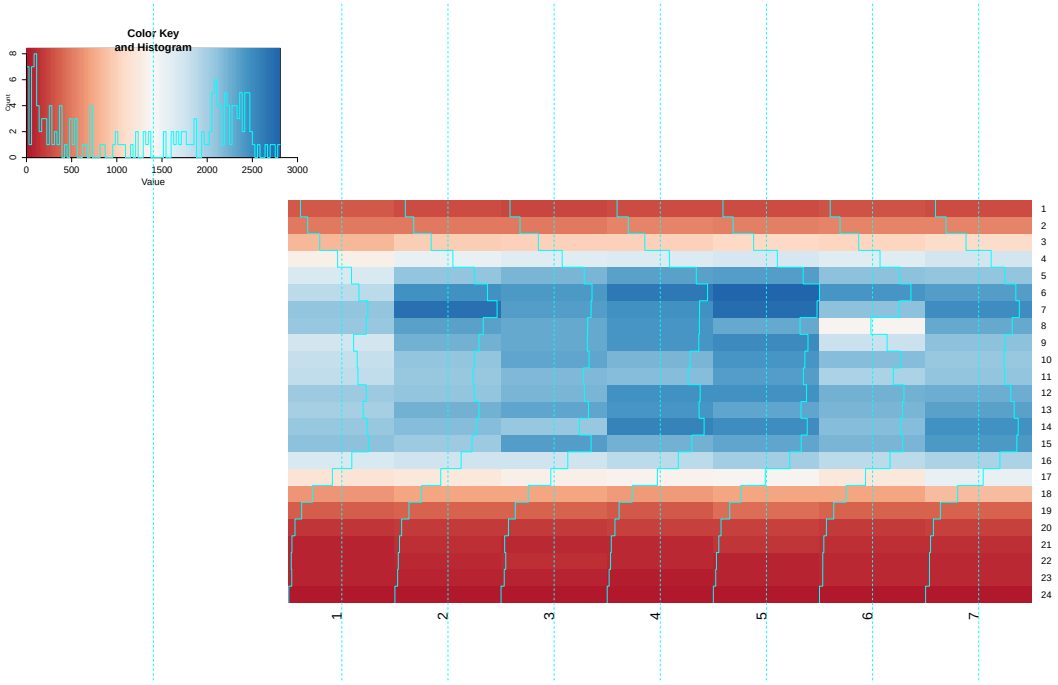


Figure 9.1.: Heatmap of all calls during the week. Values on x-axis represent days, starting from Sunday. y-axis represents hours.

We model the problem as a supervised multi-class classification problem. *Multiclass*, also known as multinomial classification is the problem of classifying instances into one of the more than two classes. In every ego profile, alters that have been called in the past are the classes. The classifier predicts at a given time and day which number(s) a user is likely to call. Each dataset and further each ego profile is evaluated independently.

9.3. Exploratory Data Analysis

By doing an exploratory data analysis, we wish to establish the rationale behind selection of dimensions (features) for predicting future calls.

We start with a simple heatmap in Figure 9.1 for the aggregated data which shows that most of the calls are made from 5*am* in the morning till 7*pm* in the evening. Most calls lasted less than one minute and very few calls were longer than ten minutes. Distribution of duration calls is plotted in Figure 9.3. Stefanis et al. (2014) argued that the safest dimensions for a call prediction system are frequency and recency of communication and observed that adding further dimensions did not improve performance of the predictive system. A quick look at the 2D plots (hours vs. weekdays) of our calls data shows that calls are concentrated in certain regions. Hence, it points to the possible relation of communication being tied to 'time-of-the-

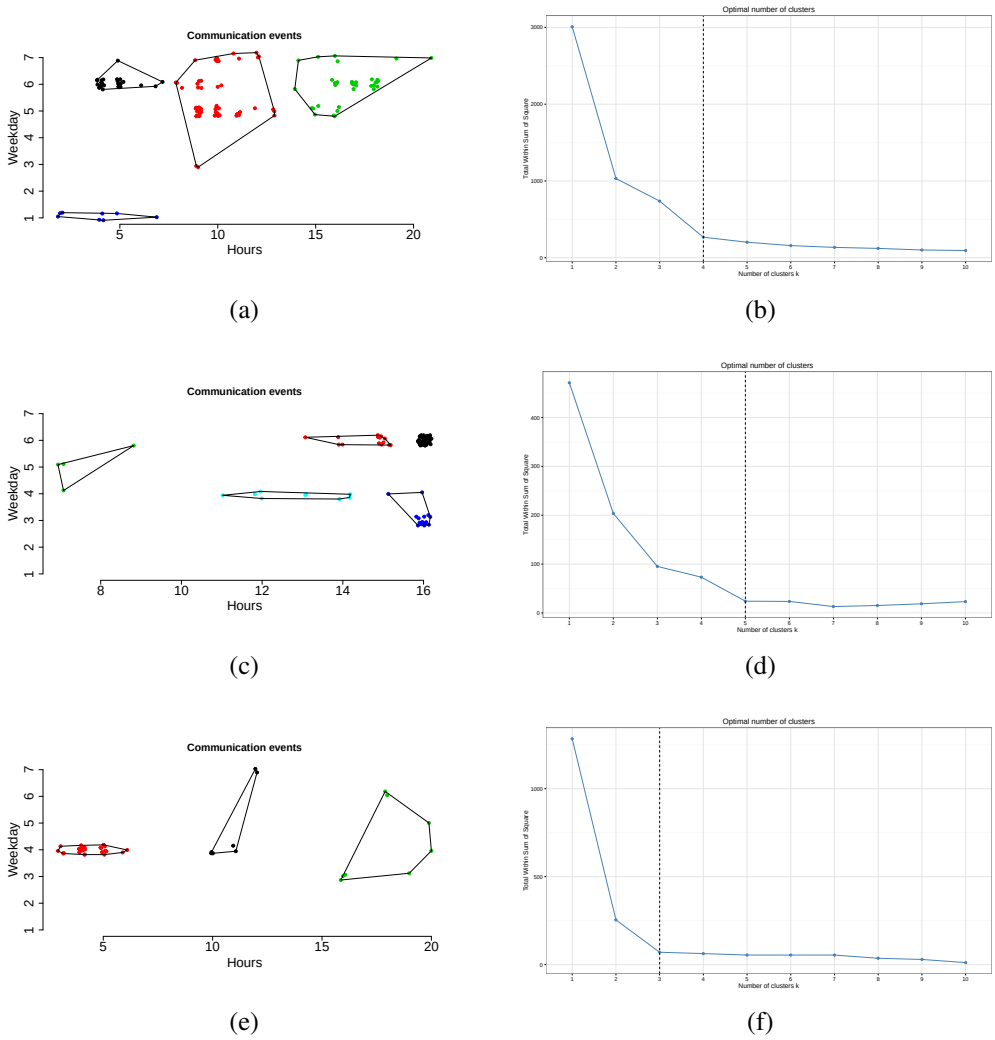


Figure 9.2.: Each plot shows a weekly and hourly calling activity for a single ego/alter pair. Figures on the left show communication events on a 2D plot. Figures on the right show the optimum number of clusters.

day'. We conducted a pretest where for each *ego-alter* pair with at least 15 communication events, we extracted the *hour of the day* and *day of the week* for each communication event and accordingly plotted it in a 2D space. We then used the *k-means clustering* algorithm to find clusters in these communications events and then found the convex hull around each of these clusters. The intuition is that small polygon like clusters indicate that the ego-alter communication has a temporal component, whereas, communication events dispersed over the plot indicate non-homogeneity in communication. Figure 9.2a shows an example of such convex hulls for three different ego alter pairs. The large cluster in red in Figure 9.2a may not show very well defined temporal calling patterns as compared to the black cluster in Figure 9.2c.

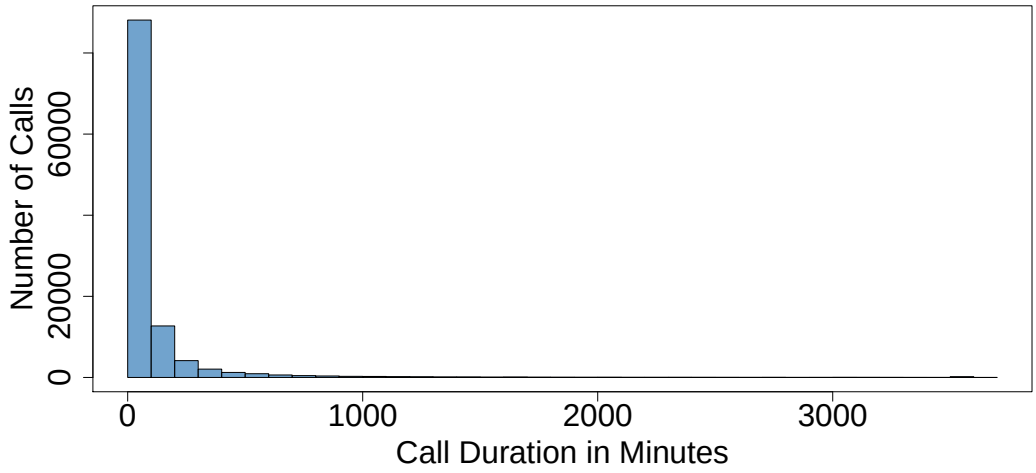


Figure 9.3.: Distribution of call duration

One issue that is inherently present with the supervised clustering schemes such as *k-means* or *DBSCAN* is to find the number of optimum clusters. At this stage the goal of the analysis is to visually observe the density of communication, therefore, one can compute clusters using any clustering algorithm of interest with different values for number of clusters. We computed within sum of square according to the number of clusters. The location of a bend (knee) in the plot (Figure 9.2b) is usually recognized as an indicator of the most appropriate number of clusters.

We next explore further statistics about the data.

9.3.1. Autocorrelation

We test the hypothesis that the majority of caller-callee interactions display temporal regularity. Formally we state the hypotheses as follows:

The alternative hypothesis (h_A) states that *the proportion of ego-alter pairs displaying periodic communication patterns is* $\geq$ *50%*. This is based on the assumption that the majority, that is, more than half, of all communication events are periodic. The null hypothesis (h_0), consequently, is that *the proportion of ego-alter pairs that display periodic communication patterns (autocorrelation) is* $<$ 50%. Autocorrelation refers to the statistical dependency between the values of a variable on related entities. In terms of time series data, like our data sets, autocorrelation implies persistence from one observation to another. Autocorrelation is a

	r_{daily}	r_{hourly}	$r_{7am-8pm}$
Reality Mining	0.55	0.89	0.65
Smartphone	0.15	0.60	0.44

Table 9.1.: Proportion of ego-alter pairs demonstrating significant autocorrelation at different time periods

common characteristic of relational and social-network datasets; since mobile calling is also a form of social interaction, it is plausible to test whether caller-callee interactions exhibit autocorrelation or not. The sample for our experiment consisted of two datasets: the call data of 783 users - henceforth referred to as *egos* - gathered using the aforementioned smartphone app (Section 8.2), henceforth called smartphone dataset; and the Reality Mining dataset from Eagle and A. Pentland (2006) collected at the Massachusetts Institute of Technology. This latter dataset comprises call and text data for 94 egos. Each dataset was analysed independently.

The data for each ego was grouped according to the contact the communication event was initiated to - an *alter* - thereby, yielding a total of 12,300 ego-alter pairs in the smartphone dataset, and 1697 pairs in the Reality Mining dataset.

The communication pattern between an ego-alter pair is modelled by a Bernoulli random variable X. As a first step, each ego-alter pair was assigned a value of one or zero based on whether the communication pattern for that particular pair was periodic or not, respectively. Our goal is to determine if the probability of success ($P(X = 1)$) is greater than 50% at the 0.05 significance level. This assignment was done first by representing communication events between the pair using a binary string, with each bit position representing a time quantum. A bit was assigned a value of one if a communication event did occur during that time quantum, and zero otherwise. We have taken into account two types of time quantum —daily and hourly.

Temporal regularity was then determined using the Ljung-Box Q test ($\alpha = 0.05$) on the string. The Ljung-Box test, also known as a *portmanteau* test, is a function of the accumulated sample autocorrelations r_k, up to any specified time lag m (Ljung and Box, 1978). As a function of m, it is determined as:

$$Q(m) = n(n+2)\sum_{k=1}^{h}\frac{r_k^2}{n-k}$$

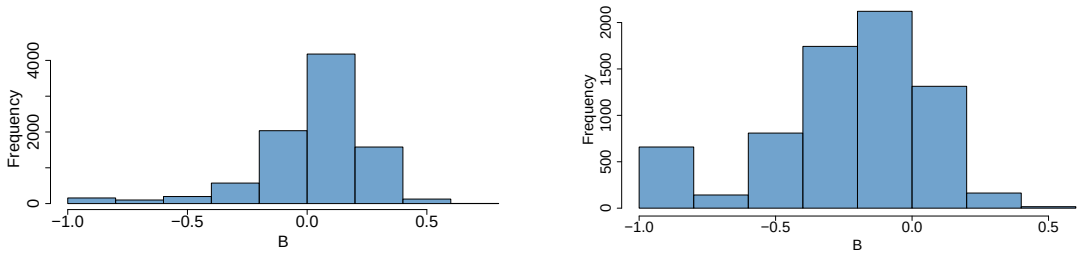
where n is the number of data points.

The null hypothesis for Ljung Box test states that the data is independently distributed i.e., any observed correlations in the data is a result of the randomness of the sampling process. The alternate hypothesis is that the data is not independently distributed; it exhibits autocorrelation (*serial correlation*). The p-value is used to decide if data points are not independently distributed. Typically, when the p-value is less than 0.05, the null hypothesis for Ljung Box test is rejected.

For each ego-alter pair, an hourly and a daily autocorrelation measure was calculated using the Ljung Box test where a $p - value < 0.05$ means there is autocorrelation. Table 9.1 lists the proportion of ego-alter pairs that displayed autocorrelation in each of the two datasets. As the hourly autocorrelation measure may have been biased by lack of activity at night hours, as one can observe in Figure 9.1, a third autocorrelation for communication events between 7am and 8pm was also calculated. Of the 12,300 total ego-alter pairs in the smartphone dataset, 4952 were filtered out during the daily autocorrelation calculation due to an insufficient number of communication events required to determine the autocorrelation.

Hypothesis test: Since the goal of the experiment was to determine whether calling behavior, in general, is periodic or not, we test the null hypothesis (h_0), which states that the proportion of ego-alter pairs that display periodic communication patterns is less than 50%. This is based on the assumption that the majority, that is, more than half, of all communication events are periodic. Based on the results for the Reality Mining dataset, we reject the null hypothesis for both the daily ($p < 2.2 \times 10^{-16}$), and the hourly autocorrelation measures ($p < 2.2 \times 10^{-16}$). This implies that more than 50% of ego-alter pairs in the Reality Mining dataset demonstrate periodic calling behaviour at the daily and daytime-hours granularity level.

For the smartphone dataset, we fail to reject the null hypothesis for the daily autocorrelation measure ($p = 1$). However, we reject the null hypothesis for the hourly ($p < 2.2 \times 10^{-16}$) autocorrelation measures.



(a)T = 1 hour

(b)T = 24 hours.

Figure 9.4.: Burstiness for the all ego-alter pairs for various values of time period T.

9.3.2. Burstiness

Several studies on human behavior indicate that the timing of communication events is characterized by long dormant periods interspersed with bursts of high activity (Barabasi, 2005; Jo et al., 2012; Y. Wu et al., 2010). Barabasi (2005) attributes this bursty non-Poisson character of human behavior to a priority-based queuing process. This view is supported by Jo et al.

(2012) who show that burstiness remains in mobile communication data even after circadian and weekly patterns have been removed, precluding the attribution of periods of inactivity to nights or weekends. They conclude that burstiness results from non-homogeneity in human task execution mechanisms.

We calculated the burstiness in the ego-alters communication based on the distribution of inter-arrival time. The burstiness parameter B is defined based on the difference of interarrival time.

$$B = \frac{\sigma - \mu}{\sigma + \mu} \tag{9.1}$$

Here σ and μ are the standard deviation and the mean of the inter-event time distribution, respectively.

The value of B can vary between $[-1, 1]$. For the most bursty behavior,$B = 1$ for homogeneous Poisson behavior $B = 0$ and $B = -1$ for completely regular behavior.

We expect that with the de-seasoning burstiness should remain as was the case in Jo et al. (2012). However, our results show a different picture. We see that the burstiness parameter shows a Poisson distribution for $T = 1$ hour in Figure 9.4a. For $T = 1$ day ($T = 24$ hours), we observe a bimodal distribution in Figure 9.4b. A significant number of ego-alter pairs exhibit regular behavior and the rest of the data has homogeneous Poisson distribution.

Burstiness parameter has been used in geology, health,communication networks to study inhomogeneous temporal patterns, primarily because of its simplicity. We observed that the burstiness parameter which is the inter-event time distribution when reduced to a single number may not be robust. When the number of events in the time series are very large only then we can expect that B would approach 1. In realistic situations we may find time series with relatively fewer events. Our view is also supported in a recent study by E.-K. Kim and Jo (2016).

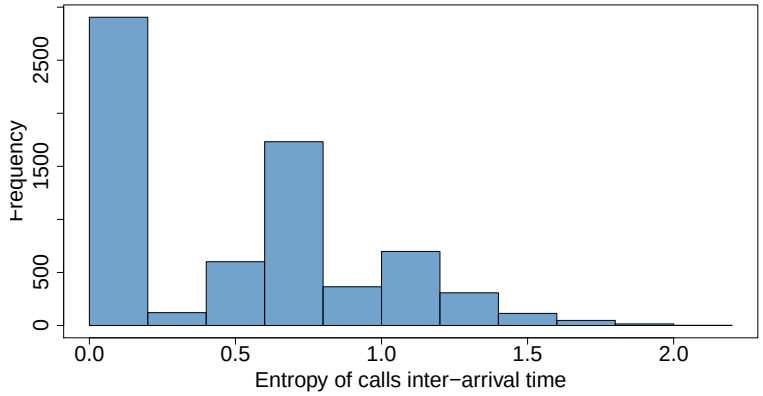


Figure 9.5.

9.3.3. Entropy

We also examined the entropy of inter-arrival time for all call pairs. Entropy is a measure of the uncertainty in a random variable. Shannon defined entropy as follows:

$$\eta = -\sum_{i=1}^{n} p(x_i) \log p(x_i) \tag{9.2}$$

Where $p(x_i)$ is the probability distribution of pairwise inter-arrival metric. When inter-arrival samples have a wider range, the entropy would be high. A higher entropy would mean there is less regularity in the calling behavior. A lower entropy points towards periodicity or burstiness. An ego-alter pair who talk around the same time everyday would have an entropy value of zero.

We divided the time into bins, each corresponding to one hour. We then created a vector of counts for each bin and calculated the entropy. Figure 9.5 shows that about 3000 ego-alter pairs have an entropy value of zero. About 45% of all ego-alter pairs have an entropy value of less than 0.5. The is consistent with the results that we obtain for the autocorrelation metric, where 44% ego-alter pairs exhibited autocorrelation in their hourly communication.

9.3.4. Recency and Frequency of Contact

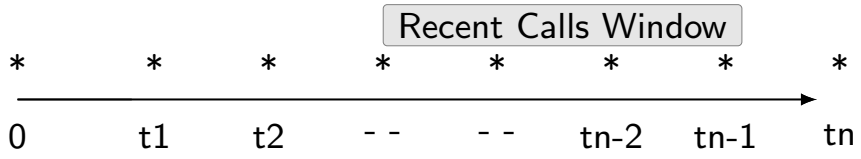


Figure 9.6.: Chronology of events

In (Nasim, Rextin, Khan, et al., 2016), our informal user study showed that about 71% of respondents either always or usually use call log to initiate a call. We hypothesized that the probability of a communication event between an ego-alter pair should be significantly less than the conditional probability of a communication event between an ego-alter pair given that there was a communication event in the near past. We checked this hypothesis by computing the conditional probability between each ego-alter pair given that there was a communication event t hours ago, where $t \in \{1,2,\cdots,12\}$. In only 3% cases the unconditional probability was higher than the conditional probability. This result supports our hypothesis. However, this conditional probability does not increase significantly and levels off very quickly as we increased t. Figure 9.6 depicts communication events in the past.

Bentley and Chen (2015) observed that most calls are to/from 5-10 contacts.This view is also supported by H. Kim, Zang, and Ma, 2013. Similarly, Bergman *et al.* Bergman et al. (2012) observed that the participants of their study did not call 47% of their contacts for 6 months. In (Nasim, Rextin, Khan, et al., 2016), we empirically found a similar but more interesting pattern; every user's call distribution very closely follow this equation: $\frac{e^a}{x^b}$, where a and b are real number that is fixed for each participant and x is the rank of the alter that varies from 1 for the alter with the most communication events and so on till the rank of the alter with the least communication events. It is worth noting that a and b both lie in a narrow range as a varied between 0-7 and b varied between 0-2.5. We observed that our equation fits the the data very well and we got a mean adjusted R^2 of 0.89 and the standard deviation was 0.16.

Plessas et al. (2016) developed a Google Android application called Calchas that aimed at predicting contacts users are most likely to call and could present the suggestions using hybrid interfaces. The algorithm behind the prediction module incorporated recency and frequency of calling. The application analyzes a recent portion of the call log and for each contact the scores for recency and frequency of communication are computed. A weighted aggregated score is then computed. The authors argued that the best results were achieved when both recency and frequency scores were weighed equally.

9.4. Feature Selection

In the previous section we conducted an exploratory data analysis in order to identify features for call prediction. A general assumption of machine learning methods is that data is *independent* and *identically distributed*. This is not true for time series data. Therefore, using machine learning methods is a disadvantage when requisite data cleansing has not been done, compared to time series forecasting techniques, in terms of accuracy. However, with careful feature selection one may obtain reasonable results.

We extracted different temporal features which we believe are most accurate in predicting the outgoing calls based on the premise that majority of the caller-callee pairs either exhibit daily or hourly temporal patterns in their interaction. For every call, the explanatory variables(features) are :time of the day (correct to the nearest minute), weekday (Sunday-Saturday), Night-call (true or false), last dialed numbers, timestamp of last dialed numbers and direction of the call (incoming, outgoing or missed). The class label is the called party's phone number. Our data is not balanced since majority of communication events take place between an ego and top contacts. We deliberately did not balance the data in order to accommodate frequency of communication in the model.

9.5. Classification

We classified our data using the Support Vector Machines (SVM) classifier, using the implementation available in R Meyer et al., 2015. We divided the data (each ego's call log) into training and test sets. We use 80% of the data (a subsequence of events) for training the model and the remaining 20% data for predicting future calls[1]. We have ensured mutual exclusivity between training and test data. We use a linear combination of the calling features for prediction. For every call in the test set, the classifier outputs probabilities against each class.

We observed that a few contacts are called more often. For our experiments we selected the mean of the calling frequency as the cut-off threshold. It is plausible to remove those callees which are very sporadically contacted. Mobile phone users have various kinds of contacts in their contacts-list such as friends, acquaintances, family members, workplace contacts, services related contacts, etc. (Bergman et al., 2012). Some of these contacts are frequently called, others are occasionally called and some contacts are not called in a long time. This view is also supported by H. Kim, Zang, and Ma, 2013. This last category of contacts, possibly does not exhibit temporal regularity at the time scale at which we are studying the problem. Hence, it is reasonable to remove them by setting a threshold on the communication frequency. In the final analysis $10,383$ caller-callee pairs are analysed in the Smartphone dataset and 1851 pairs are analyzed in the Reality Mining dataset.

9.6. Performance Analysis

Evaluation Metric

We have used the following two evaluation criteria for performance evaluation:

1. In a hypothetical situation, whenever a user presses the call button (or opens the calling interface), at time t_n, a list of contacts is displayed. Our classifier outputs probabilities for each class (contact) a user is likely to call at time t_n. These probabilities are computed and sorted and a list of contact numbers with the highest probabilities is displayed. We call this list, 'top-k recommendations'. We have also compared the performance of our approach with top-k most frequently called numbers and with last-k calls. We calculate the probability that u_i is going to call $x_j \in X_i$ (or vice versa), given that θ_j amount of time

[1]The Smartphone dataset contains data from April till September 2015. On average, the training set contained data approximately from April-August 2015 (about 16 weeks), used to predict calls made between August-September 2015 (about 5 weeks). On average, training the model took 48.35 milliseconds for each ego, on a Lenovo X1 Carbon Notebook with Intel Core i-7 CPU(2GHz) and 8GB of RAM.

has elapsed since the last communication. We denote this probability by $P(x_j|\theta_j)$. We observed that when θ_j is small or when the last communication event was a missed call from x_j (or to x_j), the probability to communicate with x_j is high. For a give θ_j, we pick the last-k' calls and include them in the results that we obtained from our classifier. We then generate a final list of most likely numbers to be dialed at any given time (within the next hour) based on the results of the classifier and last-k' calls.

2. In the second evaluation method, we measure the proportion of calls that are predicted within a certain error threshold (ε). For a given time, a 'single phone number' is predicted which the user is likely to call. We then measure how well the number is predicted with regards to different time-deviation thresholds.

Predictions are made for the users who have at least 50 communication events in the dataset. Hence we analyzed 89 users in the Reality Mining Dataset and 604 users in the Smartphone dataset. Further, we want to improve accuracy using fewer dimensions. For the last calls related features, we have used data pertinent to only the last two calls since there is a trade-off between adding dimensions to the feature set and efficiency.

Table 9.2.: Proportion of correctly predicted calls in Reality Mining dataset for various list lengths for $\varepsilon(1Hr.)$. Here last-k' are the number of last calls used in the final list.

k	last-$k' = 2$	last-$k' = 3$
1	0.44	0.44
2	0.77	0.77
3	0.78	0.77
4	0.79	0.78
5	0.80	0.78
6	0.82	0.79
7	0.84	0.80
8	0.84	0.81
9	0.85	0.82
10	0.86	0.83

Top-k recommendations

From the users' perspective, top-k recommendations should be more accurate as compared to last-k calls. We generate a list of most likely numbers to be dialed at any given time: the'top-k recommendations'. We compare the accuracy of top-k recommendations with the

Table 9.3.: Proportion of correctly predicted calls in Smartphone dataset for various list lengths for $\varepsilon(1Hr.)$. Here last-k' are the number of last calls used in the final list.

k	last-$k' = 2$	last-$k' = 3$
1	0.31	0.31
2	0.73	0.73
3	0.73	0.77
4	0.74	0.78
5	0.75	0.78
6	0.77	0.79
7	0.79	0.80
8	0.80	0.81
9	0.82	0.82
10	0.83	0.83

accuracy obtained by last-k calls. We show the performance of our approach for individual users for varying list lengths(5, 10 and 15). In Figures 9.7 and 9.8, x-axis represents the users (egos) in the dataset. For every user in the datasets, we report the accuracy achieved by top-k recommendations vs. last-k calls and top-k called numbers (most frequently called contacts). The accuracy is reported for each user in *Reality Mining dataset:* points in blue; and *Smartphone dataset:* points in red. A higher concentration of points below the identity line indicates that top-k recommendations has better performance against the respective method. Table 9.4 reports the average performance of our approach along with performance achieved by baseline methods, whereas, Table 9.2 and 9.3 report the proportion of correctly predicted calls for various list lengths.

Prediction deviation

From the service providers' perspective, accurate prediction of calls would enable them to predict users' behavior and predict periods of high usage which in turn would lead to better load balancing and, hence, better service quality. Table 9.4 reports the prediction accuracy for given deviation thresholds. These results show that a reasonable proportion of the phone calls are predictable using the proposed method. For the Reality Mining dataset 44% of the outgoing calls were predicted below one hour error threshold. For the Smart phone dataset 31% of the outgoing calls were predicted below one hour error threshold.

Table 9.4.: Average accuracy (%) with different methods.

	Reality Mining			Smartphone		
	$k = 5$	$k = 10$	$k = 15$	$k = 5$	$k = 10$	$k = 15$
Top-k called numbers	60.04	70.20	77.65	45.65	63.73	74.51
Last-k numbers	63.78	69.59	73.58	63.94	69.76	72.83
Top-k recommendations	80.70	86.66	88.46	74.96	83.72	84.08

9.7. Discussion

We analyzed whether it is possible to predict the calling behavior of mobile phone users (given the time based features), using a machine learning approach and using few dimensions. We have identified the day of the week and time as two important features which help in accurately predicting the next outgoing call. This is supported by the fact that human interaction behavior follows a circadian rhythm. We have also analyzed the situations where it is more probable that the user calls a number from one of the last called numbers.

Predictive analytics deals with understanding the data, extracting information from data and using it to predict trends and patterns. Most often the event of interest is in the future (e.g, predicting links, buying behavior, etc.), but predictive analytics can also be applied to any type of unknown event. In predictive analytics, a feature is any important piece of information about the data that might be useful for the prediction task. The purpose of a feature, other than being an attribute, would be much easier to understand in the context of a problem. Although, machine learning methods are a disadvantage when requisite data cleansing has not been done, compared to time series forecasting techniques, by careful feature selection one may obtain reasonable results.

Evaluation methods similar to ours have been used in previous studies. With a few exceptions, most previous studies used different datasets for analyzing calling behavior, therefore, a direct comparison is not equitable. Phithakkitnukoon et al. (2011) predicted the outgoing and incoming calls on Reality Mining dataset. Out of the 94 users, they selected only 30 users for experiments. The identities of those users is not disclosed in the paper, therefore, a direct comparison with their results is not possible. For completion, we have reported the performance of our method on 89 out of 94 users. The remaining 5 users had less than 50 communication events. For outgoing call prediction, they also generated a list of most likely numbers to be dialed at any given time. For the 30 random users in their experiments they achieved an accuracy of 41% if the predicted list is only allowed one entry. If the predicted list has five entries their model correctly predicted the dialed number 70% of the time. On the Reality Mining dataset we achieve an accuracy of 44% when the top-k list has one entry. Our results show more than 78%

accuracy on the Reality Mining dataset when the predicted list is allowed 5 entries. Figure 9.7 and Table 9.4 shows that our approach also performs better than the last-*k* calls on both the datasets.

Barzaiq and Loke (2011) modeled the historic call patterns of users and achieved a 35% accuracy for call prediction on a synthetic dataset. Haddad et al. (2014), reports the prediction accuracy for certain time-deviation thresholds on a dataset consisting of more than seven thousand users. Their model predicted about 17% of the outgoing calls with an error below one hour. The results for our approach, reported in Table 9.4, show 44% and 31% accuracy for Reality Mining and Smartphone datasets respectively. Haddad et al. assumes that the call arrival patterns have a Poisson distribution. An initial analysis in H. Kim, Zang, and Ma, 2013 on another large dataset suggest that the call arrival process is not Poisson for all caller-callee pairs Doran and Mendiratta, 2015. It can be argued that the particular pattern Haddad et al. mentioned in the paper was an artifact of their dataset.

In the previous models such as the ones proposed in Haddad et al., 2014 and Phithakkitnukoon et al., 2011, a baseline comparison was missing. The motivation behind our study was to come up with a method that could better predict the next call. Hence, from the user's point of view we found it imperative to check the performance of the last-*k* calls as well. It is a reasonable expectation that a call prediction approach should perform better than the current approach used for smartphone call logs i.e., displaying the recent calls in chronological order. Table 9.4 shows that our approach performs better than the last-*k* calls and most frequently called numbers list. Our call prediction approach outperformed the two baseline approaches i.e. predicting next call based on last-*k* calls and predicting next call using the most frequently called numbers' list. We found it very intriguing as it opens many exciting research questions. One of them is to see whether these results can be replicated if we take a large representative sample that can be generalized to all mobile phone users. In order to deeply understand the phone call behaviour, it is important to analyze a large call logs dataset along with other relevant information such as demographic, geographical, and socio-economic data. Another future research possibility could be an attempt to redesign the calling interface for mobile phones which could improve the user experience significantly. Such an interface, theoretically, would know the most likely people one is going to call at a given time and day. In future we would like to study how users respond to an improved call log interface.

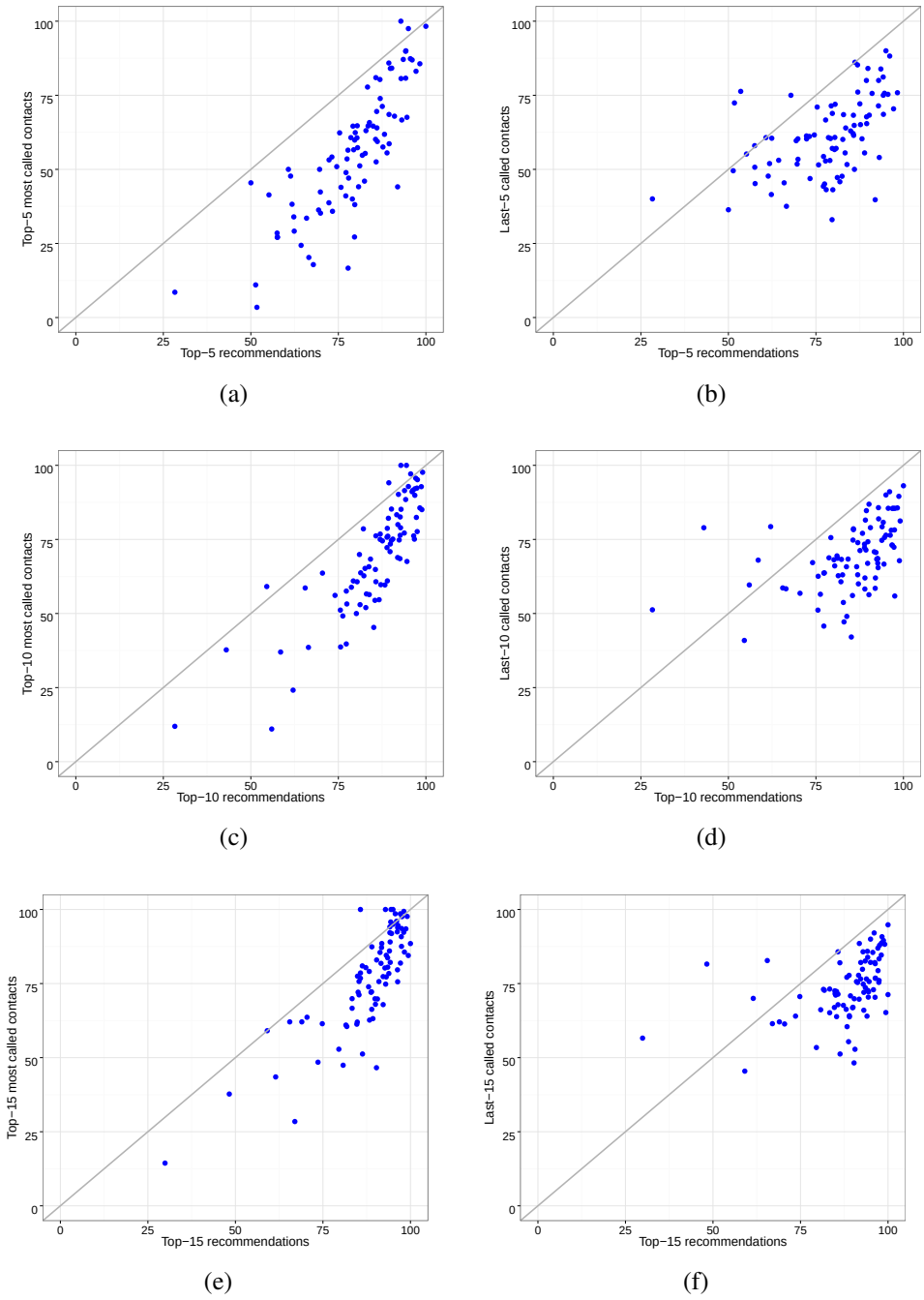


Figure 9.7.: These plots show accuracy of: top-k recommendations against top-k called numbers and last-k numbers for each user. Points below the identity line indicate that top-k recommendations has better performance against the respective baseline method. Performance is reported for: (a), (b) Reality Mining - $k = 5$. (c), (d) Reality Mining - $k = 10$. (e), (f) Reality Mining - $k = 15$.

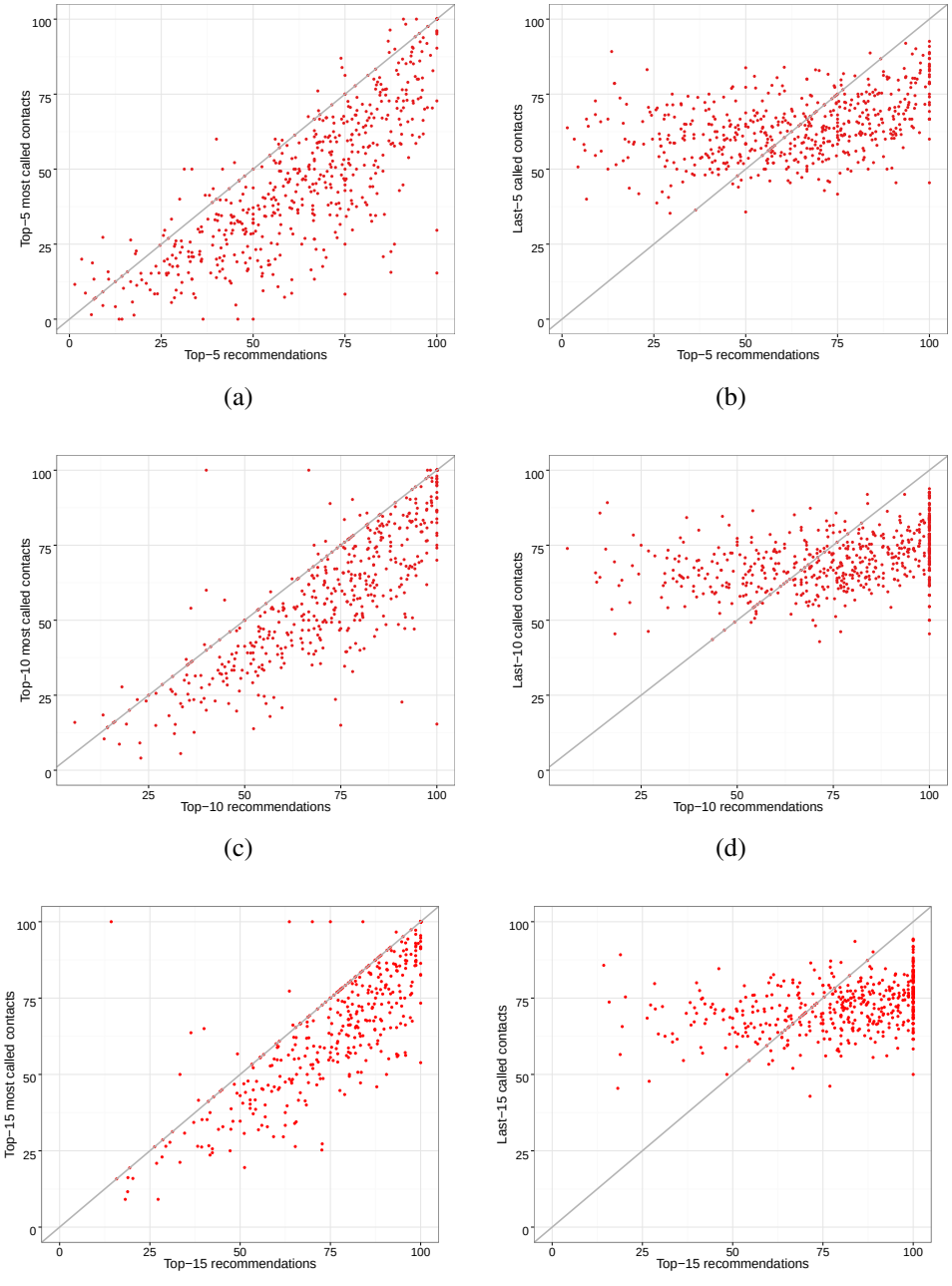


Figure 9.8.: These plots show accuracy of: top-k recommendations against top-k called numbers and last-k numbers for each user. Points below the identity line indicate that top-k recommendations has better performance against the respective baseline method. Performance is reported for: (a), (b) Smartphone - $k = 5$. (c), (d) Smartphone - $k = 10$. (e), (f) Smartphone - $k = 15$.

10. Conclusion and Future Work

This thesis was an attempt to analyze the interplay between social relations in the form of friendship ties, attributes and interaction in online social networks. In this context we analyzed composition of social circles in online social networks and showed that social circles are homophilious with respect to at least one node attribute. We showed that using the right combination of network and interaction features, links can be inferred in online covert networks. We also analyzed longitudinal dyadic interaction on cellular networks and proposed a call prediction model.

10.1. Summary of Thesis

In the first part of the thesis we showed that less persistent relations such as interaction is affected by persistent relations such as friendship ties and vice versa. We provided empirical evidence on the possibility of inferring links in online social networks using interaction information. Privacy is a concern for many users of online social networks. Our findings suggest, however, that network ties are reflected in social behavior. Privacy preserving mechanisms, i.e. hiding friends lists (Facebook) or circles (Google+), may not serve the purpose if interaction information is visible to a third party application. When there is no network information available, interactions provide substantial information about the unobserved ties.

Our study thus suggests:

- Filtering mechanisms based on homophily counteract the privacy goals expressed by hiding one's friends list.
- If users are commenting on posts because they are shown only posts written by their friends then inference of connections could be made less likely by showing users a mix of posts in their news feeds.

In the second part of the thesis we analyzed social interaction on mobile phones as a time series analysis problem. Our results imply:

- Call logs are an effective way of predicting future calls.

- A comparison of datasets from high income vs. low income countries may further deepen our understanding about usage of mobile phones in different circumstances.
- A reasonable number of ego-alter interactions exhibit temporal patterns at different time scales.

We also proposed that using temporal features, along with recency and frequency information is helpful in predicting future calls which also mitigates some of the disadvantages of machine learning approaches for our call prediction problem.

10.2. Future Work

Privacy is of immense importance to users of online social networking sites. However, in order to improve user engagement, mechanisms based on homophily counteract the privacy goals. It would be of interest to analyze the trade-offs between the quality of users' online experience and privacy settings.

Standard definition of social circles in online social networks still needs investigation. Recalling Garfinkel's account of interactions, it depicts that instead of physical co-presence, merely shared time, is relevant to identify an event as a shared event and locations as physical places are not important. Neither, the mutual friendship graph, nor the events/discussions participation is an exclusive reflection of a social circle. Online social networking sites allow users to divide their connections into groups. This is either accomplished by manual assignment or via automated assignment in a näive fashion. A better understanding of what users perceive as social circles may help in the automatic assignment of groups to online connections.

The model of user behavior assumed by current call logs is very simplistic. It supposes that the likelihood of calling a particular contact, P(c), is a monotonically decreasing function of the time elapsed since last contact. Our results show that calling behavior can be periodic and is predictable to some extent, thus calling interface can be improved. It would be useful to study how users respond to an improved call log interface. Further, it would also be of interest to explore causality and time series analysis methods for understanding relationships between network structure and network dynamics.

Appendix A

Data statistics from the survey questionnaire:

Demographic data: Out of 28 participants, 13 were males and 15 were females. **Survey Responses:** 1).Would you agree with the statement " I have contact with different people on weekends as opposed to weekdays": Strongly Agree(7%), Agree(42%), Neutral(25%), Disagree(25%). 2).Would you agree with the statement " I have experience in the usage of missed call as an indication for the other person to call back": Strongly Agree(15%), Agree(46%), Neutral(21%), Disagree(14%), Strongly Disagree(3%). 3). How often do you use call log to dial a number from your cell phone: Always(25%), Usually (46%), About 50% of the time, (21%), Rarely (7%).

Data statistics from the preliminary call log dataset:

Demographic data: Out of 13 participants, 7 were males and 6 were females. All the participants were from Islamabad Capital Territory, Pakistan. **Call Statistics:** 4682 total calls; 19% missed calls; 54% outgoing calls; 27% incoming calls and remaining 1% were rejected calls. One participant had only outgoing calls in the dataset. Data had timestamps between 01 February 2015 and 31 March 2015. **Education Level:**M.Sc (Electronics): 2; Below High School: 4; B.E (Chemical Engineering):2; M.S (System Engineering):2; High School:2; Bachelor of Arts: 1.

Bibliography

Adamic, Lada A and Eytan Adar (2003). "Friends and neighbors on the web." In: *Social networks* 25.3, pp. 211–230 (cit. on p. 72).

Agüero, Aileen and Harsha De Silva (2009). "Bottom of the pyramid expenditure patterns on mobile phone services in selected emerging Asian countries." In: *4th Communications Policy Research, South Conference, Negombo, Sri Lanka* (cit. on p. 99).

Ahn, Yong-Yeol, James P Bagrow, and Sune Lehmann (2010). "Link communities reveal multiscale complexity in networks." In: *Nature* 466.7307, pp. 761–764 (cit. on p. 12).

Aiello, Luca Maria, Alain Barrat, Rossano Schifanella, Ciro Cattuto, Benjamin Markines, and Filippo Menczer (2012). "Friendship prediction and homophily in social media." In: *ACM Transactions on the Web (TWEB)* 6.2, p. 9 (cit. on pp. 78, 84).

Airoldi, Edoardo M, David M Blei, Stephen E Fienberg, and Eric P Xing (2009). "Mixed membership stochastic blockmodels." In: *Advances in Neural Information Processing Systems*, pp. 33–40 (cit. on p. 12).

Al Hasan, Mohammad and Mohammed J Zaki (2011). "A survey of link prediction in social networks." In: *Social network data analytics*. Springer, pp. 243–275 (cit. on p. 72).

Alba, Richard D (1973). "A graph-theoretic definition of a sociometric clique†." In: *Journal of Mathematical Sociology* 3.1, pp. 113–126 (cit. on p. 9).

Aledavood, T., E López, S. G. B. Roberts, F. Reed-Tsochas, E. Moro, R. I. M. Dunbar, and J. Saramäki (2015a). "Channel-Specific Daily Patterns in Mobile Phone Communication." In: *ArXiv e-prints*. arXiv: 1507.04596 [physics.soc-ph] (cit. on pp. 97, 112).

Aledavood, T., E. López, S. G. B. Roberts, F. Reed-Tsochas, E. Moro, R. I. M. Dunbar, and J. Saramäki (2015b). "Daily rhythms in mobile telephone communication." In: *ArXiv e-prints*. arXiv: 1502.06866 [physics.soc-ph] (cit. on p. 97).

Aledavood, Talayeh, Sune Lehmann, and Jari Saramäki (2015). "On the Digital Daily Cycles of Individuals." In: *arXiv preprint arXiv:1507.08199* (cit. on pp. 97, 104).

Babbie, Earl (2010). *The practice of social research*. WADSWORTH CENGAGE Learning (cit. on p. 51).

Backstrom, Lars and Jure Leskovec (2011). "Supervised random walks: predicting and recommending links in social networks." In: *Proceedings of the fourth ACM international conference on Web search and data mining*. ACM, pp. 635–644 (cit. on pp. 74, 78, 84).

Balakrishnama, Suresh and Aravind Ganapathiraju (1998). "Linear discriminant analysis-a brief tutorial." In: *Institute for Signal and information Processing* 18 (cit. on p. 19).

Barabasi, Albert-Laszlo (2005). "The origin of bursts and heavy tails in human dynamics." In: *Nature* 435.7039, pp. 207–211 (cit. on pp. 97, 118).

Barnes, Earl R (1982). "An algorithm for partitioning the nodes of a graph." In: *SIAM Journal on Algebraic Discrete Methods* 3.4, pp. 541–550 (cit. on p. 10).

Barzaiq, Osama O and Seng W Loke (2011). "Adapting the mobile phone for task efficiency: the case of predicting outgoing calls using frequency and regularity of historical calls." In: *Personal and Ubiquitous Computing* 15.8, pp. 857–870 (cit. on pp. 96, 126).

Bastard, Irène, Dominique Cardon, Guilhem Fouetillou, Christophe Prieur, and Stephane Raux (2015). "Travail et travailleurs de la donnée." In: ed. by Lisette Calderan, Pascale Laurent, Hélène Lowinger, and Jacques Millet. Also available as `http://www.internetactu.net/2013/12/13/travail-et-travailleurs-de-la-donnee/` (cit. on p. 76).

Bauer, Eric and Ron Kohavi (1999). "An empirical comparison of voting classification algorithms: Bagging, boosting, and variants." In: *Machine learning* 36.1-2, pp. 105–139 (cit. on p. 23).

Benevenuto, F., T. Rodrigues, M. Cha, and V. Almeida (2009). "Characterizing user behavior in online social networks." In: *ACM SIGCOMM conf. on Internet measurement conference*. ACM, pp. 49–62 (cit. on pp. 2, 31, 32).

Bentley, Frank and Ying-Yu Chen (2015). "The Composition and Use of Modern Mobile Phonebooks." In: *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, pp. 2749–2758 (cit. on pp. 97, 100, 101, 108, 109, 113, 121).

Bergman, Ofer, Andreas Komninos, Dimitrios Liarokapis, and James Clarke (2012). "You never call: Demoting unused contacts on mobile phones using DMTR." In: *Personal and Ubiquitous Computing* 16.6, pp. 757–766 (cit. on pp. 101, 121, 122).

Blau, Peter M, Carolyn Beeker, and Kevin M Fitzpatrick (1984). "Intersecting social affiliations and intermarriage." In: *Social Forces*, pp. 585–606 (cit. on p. 59).

Blau, Peter Michael (1977). *Inequality and heterogeneity: A primitive theory of social structure*. New York: The Free Press (cit. on p. 71).

Bodlaender, Hans L and Babette de Fluiter (1996). "On intervalizing k-colored graphs for DNA physical mapping." In: *Discrete Applied Mathematics* 71.1, pp. 55–77 (cit. on p. 11).

Bollobás, Béla (2013). *Modern graph theory*. Vol. 184. Springer Science & Business Media (cit. on p. 9).

Bontempi, Gianluca. “Machine Learning Strategies for Time Series Prediction.” In: *Machine Learning Summer School. ULB, Brussels. Lecture* (cit. on p. 111).

Borgatti, Stephen P, Ajay Mehra, Daniel J Brass, and Giuseppe Labianca (2009). “Network analysis in the social sciences.” In: *science* 323.5916, pp. 892–895 (cit. on p. 1).

Brandes, Ulrik, M. Gaertler, and D. Wagner (2003). “Experiments on graph clustering algorithms.” In: *Algorithms-ESA 2003*, pp. 568–579 (cit. on p. 40).

Brandes, Ulrik, Garry Robins, Ann McCranie, and Stanley Wasserman (2013). “What is network science?” In: *Network Science* 1.01, pp. 1–15 (cit. on pp. 29, 104).

Brandtzæg, Petter Bae, Marika Lüders, and Jan Håvard Skjetne (2010). “Too many Facebook “friends”? Content sharing and sociability versus the need for privacy in social network sites.” In: *Intl. Journal of Human–Computer Interaction* 26.11-12, pp. 1006–1030 (cit. on p. 46).

Breiger, Ronald L (1974). “The duality of persons and groups.” In: *Social forces* 53.2, pp. 181–190 (cit. on p. 50).

Burges, Christopher JC (1998). “A tutorial on support vector machines for pattern recognition.” In: *Data mining and knowledge discovery* 2.2, pp. 121–167 (cit. on p. 22).

Burzyn, Pablo, Flavia Bonomo, and Guillermo Durán (2006). “NP-completeness results for edge modification problems.” In: *Discrete Applied Mathematics* 154.13, pp. 1824–1844 (cit. on p. 11).

Candia, Julián, Marta C González, Pu Wang, Timothy Schoenharl, Greg Madey, and Albert-László Barabási (2008). “Uncovering individual and collective human dynamics from mobile phone records.” In: *Journal of Physics A: Mathematical and Theoretical* 41.22, p. 224015 (cit. on p. 96).

Cardillo, Alessio, Giovanni Petri, Vincenzo Nicosia, Roberta Sinatra, Jesús Gómez-Gardeñes, and Vito Latora (2014). “Evolutionary dynamics of time-resolved social interactions.” In: *Physical Review E* 90.5, p. 052825 (cit. on p. 97).

Christofides, E., A. Muise, and S. Desmarais (2009). “Information disclosure and control on Facebook: Are they two sides of the same coin or two different processes?” In: *CyberPsychology & Behavior* 12.3, pp. 341–345 (cit. on p. 33).

Cover, Thomas M and Joy A Thomas (2012). *Elements of information theory*. John Wiley & Sons (cit. on pp. 80, 84).

CTIA (2015). *Annual wireless industry survey*. Tech. rep. CTIA - The wireless association (cit. on p. 112).

Cukierski, William, Benjamin Hamner, and Bo Yang (2011). “Graph-based features for supervised link prediction.” In: *Neural Networks (IJCNN), The 2011 International Joint Conference on*. IEEE, pp. 1237–1244 (cit. on pp. 73, 78, 81).

Davis, Jesse and Mark Goadrich (2006). “The relationship between Precision-Recall and ROC curves.” In: *Proceedings of the 23rd international conference on Machine learning*. ACM, pp. 233–240 (cit. on p. 83).

Derényi, Imre, Gergely Palla, and Tamás Vicsek (2005). “Clique percolation in random networks.” In: *Physical review letters* 94.16, p. 160202 (cit. on p. 11).

Domingos, Pedro (2000). “A unified bias-variance decomposition.” In: *Proceedings of 17th International Conference on Machine Learning*, pp. 231–238 (cit. on p. 23).

Domingos, Pedro (2012). “A few useful things to know about machine learning.” In: *Communications of the ACM* 55.10, pp. 78–87 (cit. on pp. 15, 23).

Doran, Derek and Veena Mendiratta (2015). “Propagation Models and Analysis for Mobile Phone Data Analytics.” In: *Propagation Phenomena in Real World Networks*. Springer, pp. 257–292 (cit. on p. 126).

Drange, Pål Grønås, Markus Sortland Dregi, Daniel Lokshtanov, and Blair D Sullivan (2015). “On the threshold of intractability.” In: *Algorithms-ESA 2015*. Springer, pp. 411–423 (cit. on p. 11).

Eagle, Nathan and Alex Pentland (2006). “Reality mining: sensing complex social systems.” In: *Personal and ubiquitous computing* 10.4, pp. 255–268 (cit. on pp. 96, 107, 109, 113, 117).

Eagle, Nathan, Alex Sandy Pentland, and David Lazer (2009). “Inferring friendship network structure by using mobile phone data.” In: *Proceedings of the National Academy of Sciences* 106.36, pp. 15274–15278 (cit. on pp. 74, 96).

Easley, D. and J. Kleinberg (2010). *Networks, crowds, and markets*. Cambridge Univ Press (cit. on p. 38).

Esling, Philippe and Carlos Agon (2012). “Time-series data mining.” In: *ACM Computing Surveys (CSUR)* 45.1, p. 12 (cit. on p. 111).

Euler, Leonhard (1741). “Solutio problematis ad geometriam situs pertinentis.” In: *Commentarii academiae scientiarum Petropolitanae* 8, pp. 128–140 (cit. on p. 9).

Facebook (2016). `http://www.facebook.com` (cit. on pp. 32, 33).

Fang, L. and K. LeFevre (2010). “Privacy wizards for social networking sites.” In: *Intl. conference on World wide web*. ACM, pp. 351–360 (cit. on pp. 45, 46).

Feld, Scott L (1981). “The focused organization of social ties.” In: *American journal of sociology*, pp. 1015–1035 (cit. on pp. 50, 71).

Fortunato, Santo (2010). “Community detection in graphs.” In: *Physics Reports* 486.3, pp. 75–174 (cit. on pp. 10, 11).

Fraley, Chris and Adrian E Raftery (2002). “Model-based clustering, discriminant analysis, and density estimation.” In: *Journal of the American statistical Association* 97.458, pp. 611–631 (cit. on p. 61).

Friedkin, N.E. (2006). *A structural theory of social influence*. Vol. 13. Cambridge University Press (cit. on p. 38).

Garfinkel (1948/2005). *Seeing sociologically*. Boulder, CO: Paradigm Press (cit. on p. 30).

Genton, Marc G (2001). "Classes of kernels for machine learning: a statistics perspective." In: *Journal of machine learning research* 2.Dec, pp. 299–312 (cit. on p. 22).

Gill, Phillipa, Martin Arlitt, Zongpeng Li, and Anirban Mahanti (2007). "Youtube traffic characterization: a view from the edge." In: *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*. ACM, pp. 15–28 (cit. on pp. 96, 104).

Girvan, M. and M. E. J. Newman (2002). "Community structure in social and biological networks." In: *Proceedings of the National Academy of Sciences* 99.12, pp. 7821–7826. DOI: 10.1073/pnas.122653799. eprint: http://www.pnas.org/content/99/12/7821.full.pdf+html. URL: http://www.pnas.org/content/99/12/7821.abstract (cit. on pp. 10, 40).

Goffman, Erving (1959). "The presentation of everyday life." In: *New York et al.: Anchor Books* (cit. on p. 45).

Goldberg, Mark, Stephen Kelley, Malik Magdon-Ismail, Konstantin Mertsalov, and Al Wallace (2010). "Finding overlapping communities in social networks." In: *Social Computing (SocialCom), 2010 IEEE Second International Conference on*. IEEE, pp. 104–113 (cit. on pp. 12, 69, 74).

Goldberg, Paul W, Martin C Golumbic, Haim Kaplan, and Ron Shamir (1995). "Four strikes against physical mapping of DNA." In: *Journal of Computational Biology* 2.1, pp. 139–152 (cit. on p. 11).

Gong, Neil Zhenqiang, Ameet Talwalkar, Lester Mackey, Ling Huang, Eui Chul Richard Shin, Emil Stefanov, Elaine Runting Shi, and Dawn Song (2014). "Joint link prediction and attribute inference using a social-attribute network." In: *ACM Transactions on Intelligent Systems and Technology (TIST)* 5.2, p. 27 (cit. on pp. 51, 73).

Good, Isidore Jacob (1950). *Probability and the Weighing of Evidence* (cit. on p. 16).

Granovetter, M. (1973). "The Strength of Weak Ties." In: *American Journal of Sociology* 78.6, pp. 1360–1380 (cit. on pp. 4, 64).

Granovetter, Mark (1985). "Economic action and social structure: The problem of embeddedness." In: *American journal of sociology*, pp. 481–510 (cit. on p. 30).

Haddad, Mohamed Ramzi, Hajer Baazaoui Zghal, Djemel Ziou, and Henda Ben Ghézala (2014). "A predictive model for recurrent consumption behavior: An application on phone calls." In: *Knowl.-Based Syst.* 64, pp. 32–43. DOI: 10.1016/j.knosys.2014.03.018. URL: http://dx.doi.org/10.1016/j.knosys.2014.03.018 (cit. on pp. 96, 126).

Hennig, Marina, Ulrik Brandes, Stephen P Borgatti, Jürgen Pfeffer, and Ines Mergel (2012). *Studying social networks: A guide to empirical research*. Campus Verlag (cit. on pp. 5, 49).

Horvát, Emöke-Ágnes, Michael Hanselmann, Fred A Hamprecht, and Katharina A Zweig (2012). "One Plus One Makes Three (for Social Networks)." In: *PloS one* 7.4, e34740 (cit. on pp. 66, 68, 72, 73).

Hosmer Jr, David W and Stanley Lemeshow (2004). *Applied logistic regression*. John Wiley & Sons (cit. on p. 16).

Hsu, Chih-Wei, Chih-Chung Chang, and Chih-Jen Lin (2003). *A practical guide to support vector classification* (cit. on p. 89).

Inc, Goolge (2010). "Adaptive contact list." Patent US 7797293 B2 (US). URL: `https://www.google.com/patents/US7797293` (cit. on p. 97).

Jamali, Salman and Huzefa Rangwala (2009). "Digging digg: Comment mining, popularity prediction, and social network analysis." In: *Web Information Systems and Mining, 2009. WISM 2009. International Conference on*. IEEE, pp. 32–38 (cit. on p. 87).

Jensen, Finn V (1996). *An introduction to Bayesian networks*. Vol. 210. UCL press London (cit. on p. 16).

Jiang, Zhi-Qiang, Wen-Jie Xie, Ming-Xia Li, Boris Podobnik, Wei-Xing Zhou, and H Eugene Stanley (2013). "Calling patterns in human communication dynamics." In: *Proceedings of the National Academy of Sciences* 110.5, pp. 1600–1605 (cit. on p. 96).

Jo, Hang-Hyun, Márton Karsai, János Kertész, and Kimmo Kaski (2012). "Circadian pattern and burstiness in mobile phone communication." In: *New Journal of Physics* 14.1, p. 013055 (cit. on pp. 97, 112, 118, 119).

Kadushin, Charles (1966). "The friends and supporters of psychotherapy: on social circles in urban life." In: *American Sociological Review*, pp. 786–802 (cit. on p. 50).

Kairam, S., M. Brzozowski, D. Huffaker, and E. Chi (2012). "Talking in circles: Selective sharing in Google+." In: *ACM annual conf. on Human Factors in Computing Systems*. ACM, pp. 1065–1074 (cit. on p. 45).

Kaltenbrunner, Andreas, Vicenç Gómez, Ayman Moghnieh, Rodrigo Meza, Josep Blat, and Vicente López (2007). "Homogeneous temporal activity patterns in a large online communication space." In: *arXiv preprint arXiv:0708.1579* (cit. on p. 96).

Kandel, D.B. (1978). "Homophily, selection, and socialization in adolescent friendships." In: *American Journal of Sociology*, pp. 427–436 (cit. on p. 38).

Kannan, R., S. Vempala, and A. Vetta (2004). "On clusterings: Good, bad and spectral." In: *Journal of the ACM (JACM)* 51.3, pp. 497–515 (cit. on p. 40).

Kernighan, Brian W and Shen Lin (1970). "An efficient heuristic procedure for partitioning graphs." In: *Bell system technical journal* 49.2, pp. 291–307 (cit. on p. 10).

Kim, Eun-Kyeong and Hang-Hyun Jo (2016). "Burstiness parameter for finite event sequences." In: *arXiv preprint arXiv:1604.01125* (cit. on p. 119).

Kim, Hayang, Hui Zang, and Xiaoli Ma (2013). "Analyzing and modeling temporal patterns of human contacts in cellular networks." In: *Computer Communications and Networks (ICCCN), 2013 22nd International Conference on*. IEEE, pp. 1–7 (cit. on pp. 97, 102, 112, 121, 122, 126).

Kotsiantis, Sotiris B, I Zaharakis, and P Pintelas (2007). *Supervised machine learning: A review of classification techniques* (cit. on pp. 15, 20, 22).

Krings, Gautier, Márton Karsai, Sebastian Bernhardsson, Vincent D Blondel, and Jari Saramäki (2012). "Effects of time window size and placement on the structure of an aggregated communication network." In: *EPJ Data Science* 1.4, pp. 1–16 (cit. on p. 96).

Lattanzi, Silvio and D Sivakumar (2009). "Affiliation networks." In: *Proceedings of the forty-first annual ACM symposium on Theory of computing*. ACM, pp. 427–434 (cit. on p. 12).

Lazarsfeld, Paul Felix, Bernard Berelson, and Hazel Gaudet (1968). "The peoples choice: how the voter makes up his mind in a presidential campaign." In: (cit. on p. 29).

Leary, Mark R (1995). *Self-presentation: Impression management and interpersonal behavior*. Brown & Benchmark Publishers (cit. on p. 45).

Lee, Conrad, Bobo Nick, Ulrik Brandes, and Pádraig Cunningham (2013). "Link prediction with social vector clocks." In: *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pp. 784–792 (cit. on pp. 74, 96).

Lenhart, Amanda (2010). *Cell phones and American adults*. Tech. rep. Pew Research Center. URL: `http://www.pewinternet.org/2010/09/02/cell-phones-and-american-adults/` (cit. on pp. 95, 112).

Leroy, Vincent, B Barla Cambazoglu, and Francesco Bonchi (2010). "Cold start link prediction." In: *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pp. 393–402 (cit. on pp. 51, 73).

Leskovec, Jure, Lars Backstrom, Ravi Kumar, and Andrew Tomkins (2008). "Microscopic evolution of social networks." In: *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pp. 462–470 (cit. on p. 12).

Leskovec, Jure, Daniel Huttenlocher, and Jon Kleinberg (2010). "Predicting positive and negative links in online social networks." In: *Proceedings of the 19th international conference on World wide web*. ACM, pp. 641–650 (cit. on p. 72).

Leskovec, Jure, Jon Kleinberg, and Christos Faloutsos (2005). "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*. ACM, pp. 177–187 (cit. on p. 12).

Leskovec, Jure and Andrej Krevl (2014). *SNAP Datasets: Stanford Large Network Dataset Collection*. `http://snap.stanford.edu/data` (cit. on pp. 39, 76).

Liben-Nowell, David and Jon Kleinberg (2007). "The link-prediction problem for social networks." In: *Journal of the American society for information science and technology* 58.7, pp. 1019–1031 (cit. on p. 72).

Lichtenwalter, Ryan N, Jake T Lussier, and Nitesh V Chawla (2010). "New perspectives and methods in link prediction." In: *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pp. 243–252 (cit. on p. 78).

Liu, Y., K.P. Gummadi, B. Krishnamurthy, and A. Mislove (2011). "Analyzing Facebook privacy settings: User expectations vs. reality." In: *ACM SIGCOMM conf. on Internet measurement conference*. ACM, pp. 61–70 (cit. on pp. 33, 46, 75).

Liu, Yunlong, Jianxin Wang, Jiong Guo, and Jianer Chen (2012). "Complexity and parameterized algorithms for Cograph Editing." In: *Theoretical Computer Science* 461, pp. 45–54 (cit. on p. 11).

Ljung, Greta M and George EP Box (1978). "On a measure of lack of fit in time series models." In: *Biometrika* 65.2, pp. 297–303 (cit. on pp. 107, 117).

Luce, R Duncan (1950). "Connectivity and generalized cliques in sociometric group structure." In: *Psychometrika* 15.2, pp. 169–190 (cit. on p. 9).

McAuley, Julian and Jure Leskovec (2012). "Learning to discover social circles in ego networks." In: *Advances in Neural Information Processing Systems 25*, pp. 548–556 (cit. on p. 51).

Mcauley, Julian and Jure Leskovec (2014). "Discovering social circles in ego networks." In: *ACM Transactions on Knowledge Discovery from Data (TKDD)* 8.1, p. 4 (cit. on p. 71).

McLachlan, Geoffrey and David Peel (2004). *Finite mixture models*. John Wiley & Sons (cit. on p. 25).

McPherson, M., L. Smith-Lovin, and J.M. Cook (2001). "Birds of a feather: Homophily in social networks." In: *Annual review of sociology*, pp. 415–444 (cit. on pp. 1, 29, 38).

Meyer, David, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel, and Friedrich Leisch (2015). *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*. R package version 1.6-7. URL: `http://CRAN.R-project.org/package=e1071` (cit. on pp. 81, 122).

Mislove, Alan, Massimiliano Marcon, Krishna P Gummadi, Peter Druschel, and Bobby Bhattacharjee (2007). "Measurement and analysis of online social networks." In: *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*. ACM, pp. 29–42 (cit. on pp. 31, 32).

Mislove, Alan, Bimal Viswanath, Krishna P Gummadi, and Peter Druschel (2010). "You are who you know: inferring user profiles in online social networks." In: *Proceedings of the third ACM international conference on Web search and data mining*. ACM, pp. 251–260 (cit. on p. 73).

mmaglobal (2015). "Mobile: The Closest You Can Get to Your Consumers." Accessed: 2016-02-05. URL: `http://www.mmaglobal.com/case-study-hub/case_studies/view/36713/` (cit. on p. 101).

Mokken, Robert J (1979). "Cliques, clubs and clans." In: *Quality & Quantity* 13.2, pp. 161–173 (cit. on p. 9).

Mondal, Mainack, Yabing Liu, Bimal Viswanath, Krishna P Gummadi, and Alan Mislove (2014). "Understanding and specifying social access control lists." In: *Symposium on Usable Privacy and Security (SOUPS)* (cit. on p. 76).

Moscovici, Serge, Carol Sherrard, and Greta Heinz (1976). *Social influence and social change*. Vol. 10. JSTOR (cit. on p. 38).

Nasim, Mehwish and Ulrik Brandes (2014). "Predicting Network Structure Using Unlabeled Interaction Information." In: *MMB & DFT 2014 Proceedings of the International Workshop SOCNET 2014*. University of Bamberg Press, pp. 57–64 (cit. on pp. 3, 87, 96).

Nasim, Mehwish, Raphaël Charbey, Christophe Prieur, and Ulrik Brandes (2016). "Investigating Link Inference in Partially Observable Networks: Friendship Ties and Interaction." In: *IEEE Transactions on Computational Social Systems* 3.3, pp. 113–119 (cit. on pp. 4, 96).

Nasim, Mehwish, Muhammad Usman Ilyas, Aimal Rextin, and Nazish Nasim (2013). "On commenting behavior of Facebook users." In: *24th ACM Conference on Hypertext and Social Media*. ACM, pp. 179–183 (cit. on pp. 3, 32, 58, 64, 74, 77, 96).

Nasim, Mehwish, Aimal Rextin, Shamaila Hayat, Numair Khan, and Muhammad Muddassir Malik (2017). "Data Analysis and Call Prediction on Dyadic Data from an Understudied Population" (cit. on p. 4).

Nasim, Mehwish, Aimal Rextin, Numair Khan, and Muhammad Muddassir Malik (2016). "Understanding Call Logs of Smartphone Users for Making Future Calls." In: *18th International Conference on Human-Computer Interaction with Mobile Devices and Services*. ACM (cit. on pp. 4, 112, 113, 120, 121).

Nastos, James (2015). "Utilizing graph classes for community detection in social and complex networks." PhD thesis. University of British Columbia (cit. on p. 10).

Nastos, James and Yong Gao (2013). "Familial groups in social networks." In: *Social Networks* 35.3, pp. 439–450 (cit. on p. 11).

Nick, Bobo, Conrad Lee, Pádraig Cunningham, and Ulrik Brandes (2013). "Simmelian backbones: amplifying hidden homophily in facebook networks." In: *Advances in Social*

Networks Analysis and Mining (ASONAM), 2013 IEEE/ACM International Conference on. IEEE, pp. 525–532 (cit. on p. 59).

Nielsen (2012). http://blog.nielsen.com/nielsenwire/social/2012/ (cit. on p. 31).

Nilsson, Nils J (1965). "Learning machines." In: (cit. on p. 16).

Nocaj, Arlind, Mark Ortmann, and Ulrik Brandes (2014). "Untangling hairballs." In: *International Symposium on Graph Drawing*. Springer, pp. 101–112 (cit. on p. 59).

O'Madadhain, Joshua, Jon Hutchins, and Padhraic Smyth (2005). "Prediction and ranking algorithms for event-based network data." In: *ACM SIGKDD Explorations Newsletter* 7.2, pp. 23–30 (cit. on p. 74).

Oulasvirta, Antti, Mika Raento, and Sauli Tiitta (2005). "ContextContacts: re-designing SmartPhone's contact book to support mobile awareness and collaboration." In: *Proceedings of the 7th international conference on Human computer interaction with mobile devices & services*. ACM, pp. 167–174 (cit. on p. 98).

Ozenc, F.K. and S.D. Farnham (2011). "Life modes in social media." In: *ACM annual conf. on Human Factors in Computing Systems*. ACM, pp. 561–570 (cit. on p. 45).

Palmore, James A (1967). *The Chicago snowball: A study of the flow and diffusion of family planning information* (cit. on p. 30).

Parthasarathy, Srinivasan, Sameep Mehta, and Soundararajan Srinivasan (2006). "Robust periodicity detection algorithms." In: *Proceedings of the 15th ACM international conference on Information and knowledge management*. ACM, pp. 874–875 (cit. on p. 107).

Peterson, Brian G. and Peter Carl (2014). "PerformanceAnalytics: Econometric tools for performance and risk analysis." In: R package version 1.4.3541. URL: http://CRAN.R-project.org/package=PerformanceAnalytics (cit. on p. 107).

Phithakkitnukoon, Santi, Ram Dantu, Rob Claxton, and Nathan Eagle (2011). "Behavior-based adaptive call predictor." In: *ACM Transactions on Autonomous and Adaptive Systems (TAAS)* 6.3, p. 21 (cit. on pp. 96, 125, 126).

Plessas, Athanasios, Vassileios Stefanis, Andreas Komninos, and John Garofalakis (2016). "Field evaluation of context aware adaptive interfaces for efficient mobile contact retrieval." In: *Pervasive and Mobile Computing* (cit. on p. 121).

Pothen, Alex (1997). "Graph partitioning algorithms with applications to scientific computing." In: *Parallel Numerical Algorithms*. Springer, pp. 323–368 (cit. on p. 10).

Potthast, Martin, Benno Stein, Fabian Loose, and Steffen Becker (2012). "Information retrieval in the commentsphere." In: *ACM Transactions on Intelligent Systems and Technology (TIST)* 3.4, p. 68 (cit. on p. 87).

PTA (2016). "Telecom Indicators." Accessed: 2016-02-08. URL: http://www.pta.gov.pk/index.php?option=com_content&view=article&id=269&Itemid=599/ (cit. on p. 99).

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: http://www.R-project.org (cit. on p. 81).

Radicchi, Filippo, Claudio Castellano, Federico Cecconi, Vittorio Loreto, and Domenico Parisi (2004). "Defining and identifying communities in networks." In: *Proceedings of the National Academy of Sciences of the United States of America* 101.9, pp. 2658–2663 (cit. on p. 10).

Raven, Bertram H (1964). *Social influence and power*. Tech. rep. DTIC Document (cit. on pp. 1, 38).

Rogers, Everett M and Dilip K Bhowmik (1970). "Homophily-heterophily: Relational concepts for communication research." In: *Public opinion quarterly* 34.4, pp. 523–538 (cit. on pp. 29, 30).

Romero, Daniel M, Chenhao Tan, and Johan Ugander (2013). "On the interplay between social and topical structure." In: *Proc. 7th International AAAI Conference on Weblogs and Social Media (ICWSM)* (cit. on pp. 69, 74).

Rosen, Kenneth T and Mitchel Resnick (1980). "The size distribution of cities: an examination of the Pareto law and primacy." In: *Journal of Urban Economics* 8.2, pp. 165–186 (cit. on p. 102).

Sarkar, Purnamrita, Deepayan Chakrabarti, and Andrew W Moore (2011). "Theoretical Justification of Popular Link Prediction Heuristics." In: *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*. Vol. 22. 3, p. 2722 (cit. on pp. 78, 84, 88).

Schlenker, Barry R (1985). "Identity and self-identification." In: *The self and social life* 65, p. 99 (cit. on p. 45).

Seidman, Stephen B (1983). "Network structure and minimum degree." In: *Social networks* 5.3, pp. 269–287 (cit. on p. 10).

Seidman, Stephen B and Brian L Foster (1978). "A graph-theoretic generalization of the clique concept." In: *Journal of Mathematical sociology* 6.1, pp. 139–154 (cit. on p. 10).

Sevtsuk, Andres and Carlo Ratti (2010). "Does urban mobility have a daily routine? Learning from the aggregate data of mobile networks." In: *Journal of Urban Technology* 17.1, pp. 41–60 (cit. on p. 96).

Simmel, Georg (1903). "The metropolis and mental life." In: *The urban sociology reader*, pp. 23–31 (cit. on p. 50).

Simmel, Georg (1908). *Soziologie. Untersuchungen über die Formen der Vergesellschaftung*. Berlin: Duncker & Humblot (cit. on pp. 50, 71).

Statista (2016). "Number of monthly active Facebook users worldwide as of 1st quarter 2016." Accessed: 2016-07-22. URL: http://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/ (cit. on p. 75).

Stefanis, Vassileios, Athanasios Plessas, Andreas Komninos, and John Garofalakis (2014). "Frequency and recency context for the management and retrieval of personal information on mobile devices." In: *Pervasive and Mobile Computing* 15, pp. 100–112 (cit. on p. 114).

Tabourier, Lionel, Anne-Sophie Libert, and Renaud Lambiotte (2016). "Predicting links in ego-networks using temporal information." In: *EPJ Data Science* 5.1, pp. 1–16 (cit. on pp. 74, 96).

Tang, Jie, Sen Wu, and Jimeng Sun (2013). "Confluence: Conformity influence in large social networks." In: *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pp. 347–355 (cit. on pp. 76, 77).

Tarde, Gabriel (1903). "The laws of imitation, trans." In: *EC Parsons. New York: Henry, Holt* (cit. on p. 29).

Tarissan, Fabien, Matthieu Latapy, and Christophe Prieur (2009). "Efficient measurement of complex networks using link queries." In: *INFOCOM Workshops 2009, IEEE*. IEEE, pp. 1–6 (cit. on pp. 78, 84).

Taskar, Ben, Ming-Fai Wong, Pieter Abbeel, and Daphne Koller (2003). "Link prediction in relational data." In: *Advances in neural information processing systems*, None (cit. on p. 73).

Thij, M ten, S Bhulai, and P Kampstra (2014). "Circadian patterns in twitter." In: *DATA ANALYTICS*, pp. 12–17 (cit. on pp. 96, 104).

Tsandilas, Theophanis et al. (2005). "An empirical assessment of adaptation techniques." In: *CHI'05 Extended Abstracts on Human Factors in Computing Systems*. ACM, pp. 2009–2012 (cit. on p. 97).

Tsaousis, Ioannis (2010). "Circadian preferences and personality traits: A meta-analysis." In: *European Journal of Personality* 24.4, pp. 356–373 (cit. on p. 107).

Twitter Glossary. https://support.twitter.com/articles/166337-the-twitter-glossary/ (cit. on p. 74).

Veropoulos, Konstantinos, Colin Campbell, Nello Cristianini, et al. (1999). "Controlling the sensitivity of support vector machines." In: *Proceedings of the international joint conference on AI*, pp. 55–60 (cit. on p. 21).

visone (2012). http://www.visone.info/ (cit. on p. 39).

Vitak, Jessica (2012). "The impact of context collapse and privacy on social network site disclosures." In: *Journal of Broadcasting & Electronic Media* 56.4, pp. 451–470 (cit. on p. 46).

Wahlke, John C (1962). *The legislative system: Explorations in legislative behavior*. Wiley (cit. on p. 30).

Walther, Joseph B (1996). "Computer-mediated communication impersonal, interpersonal, and hyperpersonal interaction." In: *Communication research* 23.1, pp. 3–43 (cit. on p. 45).

Wang, Na, Heng Xu, and Jens Grossklags (2011). "Third-party apps on Facebook: privacy and the illusion of control." In: *Proceedings of the 5th ACM symposium on computer human interaction for management of information technology*. ACM, p. 4 (cit. on p. 75).

Wasserman, Stanley and Joseph Galaskiewicz (1994). *Advances in social network analysis: Research in the social and behavioral sciences*. Vol. 171. Sage Publications (cit. on pp. 1, 49).

Wever, Rütger A (2013). *The circadian system of man: results of experiments under temporal isolation*. Springer Science & Business Media (cit. on p. 107).

Wikipedia (2016). "List of countries by number of mobile phones in use." Accessed: 2016-02-08. URL: `https://en.wikipedia.org/wiki/List_of_countries_by_number_of_mobile_phones_in_use/` (cit. on p. 99).

Wu, Xindong, Vipin Kumar, J Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J McLachlan, Angus Ng, Bing Liu, S Yu Philip, et al. (2008). "Top 10 algorithms in data mining." In: *Knowledge and information systems* 14.1, pp. 1–37 (cit. on p. 24).

Wu, Ye, Changsong Zhou, Jinghua Xiao, Jürgen Kurths, and Hans Joachim Schellnhuber (2010). "Evidence for a bimodal distribution in human communication." In: *Proceedings of the national academy of sciences* 107.44, pp. 18803–18808 (cit. on pp. 97, 118).

Xie, Jierui, Stephen Kelley, and Boleslaw K Szymanski (2013). "Overlapping community detection in networks: The state-of-the-art and comparative study." In: *ACM Computing Surveys (csur)* 45.4, p. 43 (cit. on p. 11).

Yang, Jaewon and Jure Leskovec (2012). "Community-affiliation graph model for overlapping network community detection." In: *Data Mining (ICDM), 2012 IEEE 12th International Conference on*. IEEE, pp. 1170–1175 (cit. on p. 12).

Yang, Shuang-Hong, Bo Long, Alex Smola, Narayanan Sadagopan, Zhaohui Zheng, and Hongyuan Zha (2011). "Like like alike: joint friendship and interest propagation in social networks." In: *Proceedings of the 20th international conference on World wide web*. ACM, pp. 537–546 (cit. on p. 74).

Yasseri, Taha, Robert Sumi, and János Kertész (2012). "Circadian patterns of wikipedia editorial activity: A demographic analysis." In: *PloS one* 7.1, e30091 (cit. on pp. 96, 104).

Yin, Zhijun, Manish Gupta, Tim Weninger, and Jiawei Han (2010). “A unified framework for link recommendation using random walks.” In: *Advances in Social Networks Analysis and Mining (ASONAM), 2010 International Conference on*. IEEE, pp. 152–159 (cit. on p. 73).

Zachary, Wayne W (1977). “An information flow model for conflict and fission in small groups.” In: *Journal of anthropological research*, pp. 452–473 (cit. on p. 10).

Zerubavel, Eviatar (1985). *Hidden Rhythms: schedules and calendars in social life*. Univ of California Press (cit. on pp. 96, 112).

Zhang, Zui, Hua Lin, Kun Liu, Dianshuang Wu, Guangquan Zhang, and Jie Lu (2013). “A hybrid fuzzy-based personalized recommender system for telecom products/services.” In: *Information Sciences* 235, pp. 117–129 (cit. on p. 100).

Zhao, Shanyang, Sherri Grasmuck, and Jason Martin (2008). “Identity construction on Facebook: Digital empowerment in anchored relationships.” In: *Computers in human behavior* 24.5, pp. 1816–1836 (cit. on p. 45).

Zheleva, Elena, Lise Getoor, Jennifer Golbeck, and Ugur Kuter (2010). “Using friendship ties and family circles for link prediction.” In: *Advances in Social Network Mining and Analysis*. Springer, pp. 97–113 (cit. on p. 73).