SOCIOSCOPE: HUMAN RELATIONSHIP AND BEHAVIOR ANALYSIS IN MOBILE SOCIAL NETWORKS

Huiqi Zhang, B.E., M.S.

Dissertation Prepared for the Degree of

DOCTOR OF PHILOSOPHY

UNIVERSITY OF NORTH TEXAS

August 2010

APPROVED:

Ram Dantu, Major Professor
Philip Sweany, Committee Member
Parthasarathy Guturu, Committee Member
Zuoming Wang, Committee Member
Bill Buckles, Graduate Coordinator
Ian Parberry, Chair of the Department of Computer Science and Engineering
Costas Tsatsoulis, Dean of the College of Engineering
James D. Meernik, Acting Dean of the Robert B. Toulouse School of Graduate Studies

Zhang, Huiqi. Socioscope: Human Relationship and Behavior Analysis in Mobile Social Networks. Doctor of Philosophy (Computer Science), August 2010, 133 pp., 5 tables, 40 illustrations, references, 150 titles.

The widely used mobile phone, as well as its related technologies had opened opportunities for a complete change on how people interact and build relationship across geographic and time considerations. The convenience of instant communication by mobile phones that broke the barrier of space and time is evidently the key motivational point on why such technologies so important in people's life and daily activities. Mobile phones have become the most popular communication tools.

Mobile phone technology is apparently changing our relationship to each other in our work and lives. The impact of new technologies on people's lives in social spaces gives us the chance to rethink the possibilities of technologies in social interaction. Accordingly, mobile phones are basically changing social relations in ways that are intricate to measure with any precision.

In this dissertation I propose a socioscope model for social network, relationship and human behavior analysis based on mobile phone call detail records. Because of the diversities and complexities of human social behavior, one technique cannot detect different features of human social behaviors. Therefore I use multiple probability and statistical methods for quantifying social groups, relationships and communication patterns, for predicting social tie strengths and for detecting human behavior changes and unusual consumption events. I propose a new *reciprocity index* to measure the level of reciprocity between users and their communication partners. The experimental results show that this approach is effective. Among other applications, this work is useful for homeland security, detection of unwanted calls (e.g., spam), telecommunication presence, and marketing. In my future work I plan to analyze and study the social network dynamics and evolution.

ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. Ram Dantu, for his dedication, his vision, and his support throughout the dissertation process. It has proven to be an arduous, but worthy rite of passage that has conditioned me for the road ahead.

I would like to also thank the members of the dissertation committee in taking time out of their busy schedule to read and evaluate my dissertation. Their advice and help have proven a key instrument in the completion of this dissertation work.

I thank my parents and sisters for everything that they have done for me. Especially, my parents give me unconditional support, encouragement, and belief in my endeavor.

TABLE OF CONTENTS

Page

ACKNOWLEDGMENTS iii

LIST OF TABLES vi

LIST OF FIGURES vii

Chapters

1. INTRODUCTION 1
   - 1.1 Motivation 1
   - 1.2 Contributions 3
   - 1.3 Organization of the Dissertation 4

2. BACKGROUND AND RELATED WORK 5
   - 2.1 Fundamental Concepts and Definitions in Social Network Analysis 5
   - 2.2 Fundamental Measurements in Social Network Analysis 7
     - 2.2.1 Degree Centrality 8
     - 2.2.2 Closeness Centrality 9
     - 2.2.3 Betweenness Centrality 10
     - 2.2.4 Clustering Coefficient 11
   - 2.3 Social Group Identification 13
   - 2.4 Network Evolution 16
   - 2.5 Conclusions 23

3. OVERVIEW OF THE APPROACH 24
   - 3.1 Socioscope Architecture 24
   - 3.2 Real-life Data Sets and Parameters 26
   - 3.3 Conclusions 28

4. QUANTIFYING SOCIAL GROUPS AND PREDICTING SOCIAL-TIE STRENGTHS 29
   - 4.1 Introduction 29

4.2 Dyads and Reciprocity Index 32
4.3 Social Group Identification 37
4.4 Social Tie Strength Prediction 41
4.4.1 SARIMA Prediction Model 42
4.4.2 Model Creation and Prediction Procedure 46
4.5 Experimental Results and Discussions 48
4.5.1 Quantifying Social Groups and Reciprocity index 49
4.5.2 Predicting Social-Tie Strengths 55
4.6 Conclusions 60

5. EVENT DETECTION 61
5.1 Introduction 61
5.2 Wavelet Overview 64
5.3 Wavelet Denoising 73
5.4 Change Point Detection Procedures 76
5.5 Unusual Consumption Event Detection 80
5.6 Experimental Results and Discussions 86
5.6.1 Change Point Detection 86
5.6.2 Unusual Consumption Event Detection 94
5.7 Conclusions 99

6. PATTERN RECOGNITION 101
6.1 Introduction 101
6.2 Opt-in Detection 102
6.3 Willingness Level Inference 106
6.4 Experimental Results and Discussions 108
6.4.1 Opt-in Detection 108
6.4.2 Willingness Level Inference 110
6.5 Conclusions 115

7. CONCLUSIONS 116

REFERENCES 118

LIST OF TABLES

Page

4.1 Social Groups for Phone Users 54

4.2 Prediction Errors 60

5.1 Event Dates and Locations 97

5.2 Validation results 99

6.1 Unwanted Call Rate Corresponding to the Willingness Level 114

LIST OF FIGURES

Page

3.1 The architecture of the Socioscope 25

3.2 (a) Socially close member of user 12 by feedback from UNT data; (b) Socially close member of user 70 by hand-labeling from MIT data 27

3.3 (a) Socially near member of user 12 by feedback from UNT data; (b) Socially near member of user 70 by hand-labeling from MIT data..... 28

3.4 (a) Socially far member of user 12 by feedback from UNT data; (b) Socially far member of user 70 by hand-labeling from MIT data 28

4.1 The event flow chart for phone user 29 with his partner 349..... 36

4.2 The call network of the subset of the call detail records..... 49

4.3 The affinity values for phone user 29 51

4.4 The reciprocity indices for phone user 29 52

4.5 The reciprocity indices for phone user 15 53

4.6 The reciprocity index values for phone user 39 with his partner 316..... 53

4.7 The probabilities of reciprocity for phone user 39 with his partner 316..... 54

4.8 The ACF and PACF for user 60 with his partner 2538 56

4.9 The affinity predicted and observed values for user 60 with his partner 2538 57

4.10 Residuals of the affinity values for user 60 with his partner 2538 57

4.11 The ACF and PACF for user 86 with his partner 929 58

4.12 The predicted and observed affinity values for phone user 86 with partner 929 59

4.13 Residuals of the affinity values for user 86 with his partner 929 59

5.1 Three-level wavelet decomposition 70

5.2 Three-level wavelet reconstruction 70

5.3 Harr wavelet decomposition by averaging and differencing..... 73

5.4 Harr wavelet reconstruction 73
5.5 The change points based on number of calls for simulation data..... 88
5.6 The change points based on call duration for simulation data 88
5.7 The change points based on number of calls for user3 90
5.8 The change points based on duration for user 3 90
5.9 The change points based on number of calls for user 74 92
5.10 The change points based on duration for user 74 92
5.11 Latency in the change point detection for 20 users 93
5.12 The number of incoming, outgoing and total calls per day for user 3 94
5.13 The duration of incoming, outgoing and total calls per day for user 3 95
5.14 There are events for these days for user 3 95
5.15 The number of incoming, outgoing and total calls per day for user 74 96
5.16 The duration of incoming, outgoing and total calls per day for user 74 96
5.17 There are events for these days for user 74 97
6.1 Basic service flow diagram 107
6.2 The opt-in bursts for user 1 109
6.3 The opt-in bursts for user 2 109
6.4 Willingness level and unwanted rate on Sundays..... 111
6.5 Willingness level and unwanted rate on Monday..... 112
6.6 Willingness level during 24 hours from Sunday to Saturday..... 113
6.7 Willingness level (%) vs. unwanted call rate (%) for 10 users..... 114

CHAPTER 1

INTRODUCTION

1.1. Motivation

Communication is one of the basic needs of human being and also one of the most widely practiced activities in people's daily life. Therefore, interpersonal relationships and communication are one of the most important parts in the social life of human beings. The most significant feature of a modern society is that people's daily activities are no longer limited to some particular geographical locations and their social relationships are no longer bonded to some particular persons such as kinship only. No matter how societies change, people need mutual linkages to maintain their interpersonal relationship. As modern technology advances, people can overcome the spatial distance and the constraint of time to maintain social relationships by telecommunication and social media. The telecommunication technologies and social media serve as a bridge between one person and another. People use modern technology to reconstruct new forms of social relationships. Since modern telecommunication technology advances rapidly, styles and means of interpersonal communication changes correspondingly. It results in a lot of changes in personal intimate relationships.

Modern telecommunication and Internet technologies such as mobile communications and social media make people world wide form a wide area social network (WASN). Mobile phone plays a significant role in everyone's daily life. It has become the most popular communication tool. The individual-to-individual connection in the mobile age would substitute place-to-place connection in the fixed

telephone age in the near future. Since the usage of mobile phone could shorten the physical distance, save time, money and physical strength, it becomes more convenient for people to communicate with others and accelerates information exchange among interpersonal contacts. As a result, it also accelerates changes and evolution of social relationships. We may communicate with our family members, relatives, friends and colleagues, and find new friends who have the same interests like ours at any time and almost anywhere. In WASN people form many different groups or clusters based on interests, goals, etc. Since mobile phone becomes an important tool of human daily life in modern society, different telecommunication patterns may reflect different human relationships and behaviors, and the telecommunication pattern changes of human may indicate some signs of social relationship and behavior changes. For example, the calling patterns of a person with his/her friends are different from those of his/her with spammers.

Organizations or individuals may be interested in different social network properties. For example, people in homeland security related departments are interested in special groups of people such as terrorists, robbers. Businessmen want to know potential consumers who are interested in their products. Network designers and operators want to know overall distributions and patterns of users in order to efficiently and effectively use and distribute resources and enhance quality of services. Almost all existing social network research has focused on overall social network structures and properties such as clusters, communities. There is a lack of one-to-one or one-to-many relationships and behaviors measurement in more details in special groups or clusters of people. These detailed features of people's relationships are more important for detecting terrorists, spammers, etc. Because of the diversities and

complexities of human social behavior, one technique only cannot detect different features of human social behaviors. Therefore I have used multiple probability and statistical methods, and integrate them for social network and human behavior analysis from macro to micro level.

1.2. Contributions

The main contributions of this dissertation are as following:

- Proposed a socioscope model for analyzing the properties of mobile social network structures and human behavior, quantifying measurement and predicting interpersonal relationships, detecting events and recognizing calling patterns based on human telecommunication patterns.
- Developed and evaluated innovative approaches for group identification
    - Proposed a new reciprocity index for quantifying reciprocity property of human telecommunication.
    - Proposed an affinity model in which the new reciprocity index was integrated for quantifying interpersonal relationships.
- Map call-log data into time series of social tie strengths by the affinity model so that social tie strengths are functions of time, and predict social tie strengths by seasonal auto regressive integrated moving average (*SARIMA*) model.
- Combine wavelet denoising and sequential detection methods for detecting change points.
- Proposed the inhomogeneous Poisson model for detecting unusual consumption events.
- Proposed the dynamic size window model combined with exponentially weighted moving average (EWMA) method for detecting opt-in bursts.

- Proposed a Bayesian inference model for quantifying willingness level to communicate with each other based on human telecommunication patterns.

The socioscope consists of several components including zoom, scale and analysis tools which are used for analyzing network structures, discovering social groups, quantifying and predicting relationships, detecting events, and recognizing calling patterns. It is extensible and new tools can be added as needed. By zooming in multiple scales can be used to analyze social group members' behavior up to one-to-one. By zooming out general social network structures and properties can be analyzed.

1.3. Organization of the Dissertation

The remainder of this paper is organized as following: In Chapter 2, I briefly review the background and the related work. Chapter 3 presents the model, architecture, components, and dataset for the socioscope. I describe the methods for identifying social groups, quantifying reciprocity index, and predicting social-tie strength, perform experiments, conduct validation, and discuss the results in Chapter 4. I describe the approaches for detecting change points and unusual consumption events, perform experiments, conduct validation, and discuss the results in Chapter 5. I describe the methods for detecting opt-in calling patterns and quantifying willingness levels, perform experiments, conduct validation, and discuss the results in Chapter 6. Finally, I present the conclusion in Chapter 7.

CHAPTER 2

BACKGOUND AND RELATED WORK

Study of social networks has been applied to modern sociology for some time. The major applications focus on measuring interpersonal relations in groups, describing properties of social structures and individual social environments (Wasserman and Faust, 1994, 5).

Social networks are defined as the set of actors (individuals) and the ties (relationships) among them (Wasserman and Faust, 1994, 7). These relational ties and actors compose the fundamental interests in social networks. Also, social groups can be defined as a set of people who have common interests, like the same subjects, share their experience, express similar way of thinking and reacting, share the same opinions, do similar things, and have the same goals. They actively exchange information. Facing new circumstances, they discuss with each other to decide what to do.

2.1. Fundamental Concepts and Definitions in Social Network Analysis

Social networks are usually represented and analyzed via a graph. A graph can be presented as $G=(V,E)$ where V is the set of elements called vertices (nodes or actors) and E is the set of unordered pairs of vertices called edges (links or ties).

A vertex i is adjacent to j if between i and j there is an edge included in the set E. This edge is denoted as e_{ij}. The vertices i and j are called endpoints of e_{ij}, and the edge e_{ij} is said incident with vertices i and j. Edges in networks can have directions.

If edges have directions, a graph is called directed graph or digraph. Otherwise, it is called undirected graph. All edges in an undirected graph are symmetric.

Degree of a vertex *i* is defined as the number of edges incident with *j*. The maximum degree of graph *G* is the largest degree over all vertices. In a directed graph, vertices may have two types of degrees, in-degree and out-degree. In-degree is the number of edges pointing to vertex *i*. Out-degree is the number of edges leaving from the vertex *i*.

Graphs can be of two types based on the number of relations characterized by an edge between two vertices. First, an edge could have one relationship, i.e., whether or not that relationship is present between those two vertices. This kind of relationship is called simple relationship and the graphs having only simple relationships are called simplex graphs. Second, an edge could have a label that has a vector of multiple features that characterize the relationship between those two vertices. These relations are called multiplex relations and the graphs with only these kinds of relationships are called multiplex graphs.

Graph vertices and edges contain information. When this information is a simple label the graph is called a labeled graph. In other more complex representations, vertices and edges may contain some more information like vertex and edge attributes (or weighted). In this case the graph is called an attributed (or weighted) graph. An attributed (or weighted) graph is further distinguished between vertex-attributed (or weighted) and edge-attributed graphs.

For an edge-weighted graph, each edge e_{ij} can be assigned some value w_{ij}. A weighted graph can be represented as $G=(V, E, W)$ where W is the set of edge weights. A path connecting vertices *i* and *j* is defined as an alternating sequence of

vertices and edges. Path length is the sum of all the weights of the edges belonging to this path. The geodesic path between two vertices *i* and *j* is the path with the shortest distance, and this distance is called geodesic distance, and denoted as d_{ij} .

A graph *G'*= (*V', E'*) is a sub-graph of *G*= (*V, E*) if its vertex set *V'* and edge set *E'* are subsets of *V* and *E* respectively. It can be written as *G' = (V', E')*, where $V' \subseteq V$, and $E' \subseteq E$. *G* is called a super-graph of *G'*. This sub-graph is called an induced sub-graph of *G* if for every pair of vertex *i* and *j* of *G'*, e_{ij} exists in *G'* if and only if there is an edge e_{ij} in *G*. In a sub-graph, a boundary vertex is defined as a vertex which has connections with vertices belonging to other sub-graphs. The boundary size of a sub-graph *G'* is defined as the number of its boundary vertices contained.

A clique in a graph is a set of pairwise adjacent vertices, or an induced subgraph which is a fully connected graph. The clique problem is the problem of determining whether a graph contains a clique of at least a given size *k*. The corresponding problem of finding the largest clique in a graph is called the maximum clique problem.

The number of vertices or nodes within a graph denotes the order or size of the graph. The density of the graph is the ratio of the number of edges to the theoretical maximum number of edges possible. The maximum number of edges is *n(n − 1)/2* for a undirected graph and *n(n−1)* for a digraph *n* vertices.

2.2. Fundamental Measurements in Social Network Analysis

Centrality is one of the most important and frequently used measurements in social network analysis (Carrington, 2005). It is a descriptive characteristic for actors

or groups of actors with various structural properties and a crucial parameter for understanding and analyzing actor's roles in social networks (Newman, 2005). Centrality has diverse definitions because of different understandings of social power and various applications (Carrington, 2005). The most widely accepted definitions of centrality are proposed in by Freeman (Freeman, 1979). In these definitions, centrality is measured mainly based on three aspects, degree, closeness, and betweenness. These measurements offer a sense of how important each actor is in the network. They focus on the actor and are basics for network modeling because they indicate the volume or strength of ties flowing from and to each actor. Using these measurements of centrality, I was able to uncover particularly important actors within the network that are influential in the spread and activation throughout the network. The importance of certain actors is indicated based on their location within a network relative to other actors.

2.2.1 Degree Centrality

Degree centrality is defined as the number of ties which are incidents with a given actor. This measurement usually reflects the popularity and relational activity of an actor (Marsden, 2002; Frank, 2002; Newman, 2005). The degree centrality of an actor in a network reflects the number of other actors to which this actor is related. An actor with a high degree of centrality is described as a characteristic highly connected with many others. Degree centrality is defined as

$$C_D(i) = \sum_{j=1}^{n} e_{ij}$$

where n is the number of actors in the network and $e_{ij} = 1$ if i and j are connected by an edge, $e_{ij} = 0$, otherwise.

Freeman (1979) proposed a relative measure of degree centrality in which the actual number of connections is related to the maximum possible number that could exist. The relative degree centrality is defined as

$$C_D^{'}(i) = \frac{1}{n-1}\sum_{j=1}^{n} e_{ij}$$

Degree centrality measures network activity. A high degree centrality shows its high impact in the network. Therefore, a high degree centrality represents core actors. Actors with high degree centrality have been shown to be more active (Frivolt and Bielikov, 2005) and influential (Memon et al., 2008). The distribution of degree centrality scores can be used to describe the characteristics of online communities (Fisher, 2005) and to visualize the roles of members (Welser et al., 2007) within subgroups.

An actor with a high degree centrality should be recognized as a major transfer point of information flows. Since it has many alternatives, such an actor is relatively independent of the control of others, and resistant to information delay or distortion; however, there is a problem of information overload or conflict. This measure tells nothing about the knowledge or information the actor has.

2.2.2 Closeness Centrality

Closeness centrality measures how close an actor is to all other actors. A high closeness centrality indicates independence from the control potential of other actors in the network or of the flow of activation from others. Closeness is a measurement that defines an actor as central according to the extent to which the actor is close enough to many others to activate them directly. Therefore, an actor needs to not rely on others to spread activation. If all actors in a network are close to each other,

activation originated from any actor will flow through the rest of the network quickly; thus, high closeness centralities are associated with short time of activating the schema and low costs depending on the content of relational ties being modeled (Freeman, 1979). In principle, actors with high closeness centrality should be able to connect more efficiently or easily with other actors, making them more likely to participate in subgroups. Closeness centrality is defined as

$$C_c(i) = [\sum_{j=1}^{n} d_{ij}]^{-1}$$

where d_{ij} is the geodesic distance from actor *i* to actor *j* and the geodesic distance is the number of links in the shortest possible path from one actor to another.

Comparison of closeness centrality must be done in networks of the same size. To solve this limitation, Beauchamp (1965) suggested a relative definition of closeness centrality which is

$$C_C'(i) = [\frac{1}{n-1} \sum_{j=1}^{n} d_{ij})]^{-1}$$

Closeness centrality has been used to identify important actors within social networks (Crucitti, Latora, and Porta, 2006; Kurdia, 2007; Ma and Zeng, 2003) and identify members with a strong sense of community.

An actor with high closeness centrality can interact with all others at low resource cost and information distortion. If the length of path is defined by the number of direct ties it contains, an actor with high closeness would be an excellent monitor of the information flow in the network.

2.2.3. Betweenness Centrality

The shortest path that links two vertices, *i* and *j*, in a network is a geodesic. Any point that falls on a geodesic links a pair of points standing between these two

points. Betweenness centrality is an indicator of control within a network. Furthermore for semantic networks a node with a high betweenness centrality has a higher probability of getting activated or of activating other nodes. The betweenness measure is defined in terms of probabilities; in cases where there is more than one possible path, it considers the probability of using a particular path (Freeman, 1979). Betweenness centrality is computed as

$$C_B(k) = \sum_{i=1}^{n} \sum_{j=1}^{n} g_{ij}(k) / g_{ij}$$

for all $i < j \neq k$, where g_{ij} is the number of geodesic paths from *i* to *j* that contain *k* and $g_{ij}(k)$ is the number of geodesic paths that pass through *k*. Therefore, $g_{ij}(k) / g_{ij}$ is the probability that *k* falls on a randomly selected geodesic connecting *i* and *j*. Betweenness centrality reflects the likelihood that a node will be activated as associations spread throughout the network. A node being on many paths between other pairs of nodes is between many nodes and will have a high betweenness centrality (Henderson et al., 1998).

An actor with high betweenness centrality also controls the channel resources of information transfer and hence tends to be attacked, but it may not control the knowledge or information resources. It can play the role of a coordinator or a third party in information transfer. The change in these actors may influence the information transfer between the source and the recipient (for example, it may have to choose another path). This measure is useful in the assessment of the efficiency and control ability of channels. This measure is a sensitive indicator of important changes in the network topology.

2.2.4 Clustering Coefficient

Clustering coefficient is a local property measuring the connectedness of the neighbors of the vertex. The clustering coefficient of a vertex is also referred as the density of its open neighborhood, which represents how close each vertex's neighborhood is to a fully connected clique. The clustering coefficient of a vertex can be calculated as the ratio of the number of edges between the neighbors of the vertex and the number of all possible edges between the neighbors.

Let E_i denotes the number of edges between the vertex *i* within its neighbors. Suppose vertex *i* in the network has k_i neighbors. There could exist the maximum possible $k_i(k_i-1)$ edges among the vertices within the neighborhood for directed graphs. The local clustering coefficient C_i for directed graphs is defined as

$$C_i = \frac{E_i}{k_i(k_i-1)}$$

For undirected graphs there could exist the maximum possible $k_i(k_i-1)/2$ edges among the vertices within the neighborhood. The local clustering coefficient C_i for undirected graphs is defined as

$$C_i = \frac{2E_i}{k_i(k_i-1)}$$

The clustering coefficient of the network can be taken to be the average of the clustering coefficients of all the vertices for the whole network. Watts and Strogatz (1998) define a clustering coefficient of all the vertices *n* for the whole network as

$$\overline{C} = \frac{1}{n}\sum_{i=1}^{n}\frac{E_i}{k_i(k_i-1)}$$

for directed graphs and

$$\overline{C} = \frac{1}{n}\sum_{i=1}^{n} \frac{2E_i}{k_i(k_i - 1)}$$

for undirected graphs.

The clustering coefficient is the average of the individual clustering coefficients. The weighted overall clustering coefficient is the weighted mean of the clustering coefficient of all the vertices with each one weighted by its degree. The cluster coefficient tends to 1 if most of the partners of each actor are directly related. On the contrary, the clustering coefficient tends to 0 if the network is hierarchical and the partners of each actor are not related. Clustering coefficients are often applied to detect small-world networks and the degree of hierarchy of local relational structures. At the network level, the degree of hierarchy of local relational structures is called transitivity. If an actor's ego-network is a clique, meaning the absence of hierarchy, the actor and its partners all have equal structural power and the network as a whole becomes more transitive.

Centralization is an expression of how tightly the network is organized around its most central actors (Freeman, 1979).

2.3 Social Group Identification

A set of nodes can form a group or community if they are more closely related to one another than the rest of the network. Such nodes connect with a higher density within the group and are very sparsely connected to the rest of the network.

Brass (1995) summarized the social network measures involving the relational ties, the actors, and the overall consequences of the network topology. Carley et al. (2002) proposed a concept meta-matrix describing a composite network that

incorporates the multi-dimensionality of interpersonal relations. The meta-matrix is based on that network dynamics are functions of (1) the social structure, (2) the distribution of knowledge and information, (3) the interrelations between domains of knowledge, and (4) the distribution of work and requirements. These aspects of an organization in the meta-matrix construct serves as input into an agent-based network simulation, which evaluates the organization's ability to perform tasks, communicate effectively, and so on. By the Loom system in Donath et al. (1999) users can visualize social relationships on the basis of Usenet conversation. Sack (2000) used the several text analysis procedures to compute a net and visualizations of social relations among authors, some discussion themes frequently used in the conversations and the semantic networks which represent the main terms in the discussions and relations in them. Microsoft Netscan (2006) was used for searching, visualizing, and analyzing newsgroups. The various visualization techniques and extensive relation analyses and content analyses are combined to allow for an improved navigation in the Usenet. The Friend-of-a-Friend (FOAF) project (Brickley and Miller 2004) explored the application of semantic web technologies to describe personal information including professional and personal lives, their friends, interests and other social relations. Dill et al. (2001) investigated an extensive characterization of the graph structure of the web with various features and consider the subgraph of the web consisting of all pages containing these features.

Flake et al. (2000) proposed a network flow based approach to partitioning the graph into communities. Broder et al. (2000) used block technique which focuses on the pattern of connectivity to find clusters of relationships that might be hidden. Ogatha (2001) discussed the strength of social relations between two persons

measured with the email conversation. The relation is strong if email between two persons is exchanged frequently, recently and reciprocally and a formula was used for the strength, which is a function of user determined importance weights and the number of received and sent emails. Newman and Girvan (2002, 2006) investigated the automatic detection methods of subgroup structure of social networks based on expected density and edge betweenness. Flake et al. (2002, 2004) and Hopcroft et al. (2003) focused on identifying tightly-connected clusters or communities within a given graph and inferring potential communities in a network based on density of linkage. Viegas and Smith (2004) used the self-identified nature of the online communities to study the relationship between different newsgroups on Usenet.

To determine the best community structure by finding the optimal modularity value, Girvan and Newman (2002) proposed an algorithm by removing edges with high betweenness centrality. By repeated calculation of the edge betweenness and removal of edges the entire network is decomposed into its constituent communities. Ravasz et al. (2002) proposed an approach by optimizing modularity via topological overlap. Newman (2004) proposed a fast approximation to compute betweenness centrality for community detection. Betweenness centrality is a measure of the number of times a node is on the shortest path route amongst all other pairs of nodes. Newman et al. (2004) defined a modularity function for the quality of any clustering or community detection algorithm. The modularity function measures the fraction of all the edges that connect within the community to the fraction of edges that are across communities. Clauset et al. (2004) presented greedy heuristic method and Newman (2004) used efficient update rules for merging clusters during the hierarchical clustering process for community detection.

Doreian, et al. (2005) proposed the generalized block modeling method to enable partitions of smaller, more precise block types and predefined block structures based on attribute data. Lin et al. (2006) identified a group of blogs which are mutually aware of each other. Each pair of blogs was weighted with an association score based on the different actions such as post-to-post links, comments, trackbacks that indicate awareness between the corresponding blogs.

Chi et al. (2007) applied spectral clustering methods based on the analysis of eigenvectors of a graph to community detection. Newman (2006) showed that spectral clustering methods can affect the optimization of the modularity score.

Another approach for community detection is the kernel method (Taylor and Cristianini, 2004). A kernel provides a simple mapping from the original, high dimensional feature space to a set of correlation scores between the data points. The algorithm of Girvan and Newman (2002) is the computation of all pairs of shortest paths, while in kernel methods it is the calculation of an inverse.

2.4 Network Evolution

The understanding of how networks evolve has been a topic of interest to social network researchers. These researchers realize that relational behavior and network structure are intertwined. An actor's relational strategy depends to some extent on the structure of previous relationships. At the same time, the actor's new relationships contribute to a changed network structure that again influences its actions (Gulati, 1998). Network evolution principles are mainly based on complex network theory. The process of link formation is based on certain rules of attachment. Network change consists of changes in the number of actors and patterns of link formation (Palla, 2007). Structural network change is a form of network change

whereby new linkages are formed with new partners. Studies on new partner search in networks have broadly focused on two issues. One issue is about distribution of ties among actors in a network. In many real world networks the distribution of linkages among actors is highly unequal (Dorogovtsev and Mendez, 2003; Barabasi et al., 2002). Barabasi (2000) showed how actors accumulate new linkages in proportion to the number of linkages they already have. Following this rich get richer principle of growth, the resulting network structure consists of a few highly connected actors called stars in combination and many weakly connected peripheral players. The second issue is that new partner search concerns the process of local link formation and the process of distant link formation.

Local link formation implies that new partners are found through an actor's existing network which is called an ego network, and that the new partner is already known to other partners in the neighborhood. The overall network structure resulting from local link formation is a network composed of dense cliques of actors, which indicates that they are highly connected to each other. Local link formation of an actor and the degree of clique formation in a network can be measured by calculating the clustering coefficient.

The first network study that combined local and distant link formation originate from complex network studies. Watts and Strogatz (1998) modeled the process of local link formation and found that, with the addition of just a handful of distant linkages, a specific network structure is generated, which they called a small world. This means that although large networks have relatively few linkages compared to the number of actors, the reach is higher than expected (Newman, 2001). While solely local link formation results in dense cliques of connected actors, the

average distance to reach all actors in a network is very large. The distance between two actors is indicated by the number of other actors one has to surpass in order to reach the other. Watts and Strogatz (1998) found that the average distance between all actors in a network is sharply reduced when a relatively small number of distant linkages which is referred to as random linkages are added to the network that serve as shortcuts between these local cliques. Jackson and Rogers (2006) focused on link formation with new partners in social networks. They found that large social networks evolve into small worlds, because people meet friends of friends and strangers. The process of link formation is generated by an algorithm that makes actors form both local linkages and distant linkages, which implies that distant link formation resembles the meeting of strangers.

Since most real-world networks are temporal and dynamic in nature. Communities may merge to form a larger community or a single community may split into a number of smaller groups.

Desikan and Srivastava (2004) investigated the change in metrics of a set of web pages over time for the graph as a whole and for single nodes. They found that temporal metrics, such as their page usage popularity, can be effectively used to boost ranks of recent popular pages to those that are more obsolete.

Kumar et al. (2003) have studied the evolution of the blog graph and find that there was an emergence of stronger community structure at a microscopic level. Shi et al. (2007) investigated the blog datasets and found that there are a lot of similarities between the blog graphs and the Web graphs by comparing them to the known parameters of the Web. Results showed power-law slopes of about 2 to 2.5 which are close to 2.1 slope obtained in the Web. Kumar et al. (2003) obtained similar values in

their study. Leskovec, Kleinberg and Faloutsos (2005) studied the trends of social network evolution and found that over time graphs increase in density and the average distance between nodes decrease. They proposed the "Forest Fire" model to explain the growth and evolution of dynamic social network graphs. In this model, new nodes arrive one at a time and attach themselves to an existing node preferentially. Once the node is connected, it performs a random walk in the neighborhood and creates new links locally. The process is then repeated for each of the new nodes that are linked to during the random walk. They found that the distribution of the sizes of such cascades also follow a power law distribution.

In the papers of Borgs et al. (2004), Holme and Newman (2006), Sarkar and Moore (2005) social network evolution was studied as its members' attributes change. Backstrom et al. (2006) performed a large-scale empirical analysis of social network evolution in which interactions between people are inferred from time-stamped e-mail headers. in Kossinets and Watts (2006) proposed the social network evolution model of topics over time.

Koka et al. (2006) have combined multiple indicators of relational behavior into four different types of network change. The network can expand, churn, strengthen or shrink. Each network change is brought about by a specific combination of changes in tie creation, tie deletion, and by changes in an actor's number of links and number of partners.

While Koka et al. (2006) presented four types of network change they found that only an expanding network and a churning network are reflections of structural change, because new alliances are formed with new partners. An expanding network is brought about by an increase in new alliances without deletion of old alliances,

together with an increase of more different partners. A churning network reflects the formation of new alliances and the deletion of existing alliances. While the average number of partners remains stable, there is an increasing variety in the identity of partners.

While changes in the number of linkages (tie creation or deletion) and changes in the number and identity of partners already provide important insights into structural changes in the network, Jackson and Rogers (2006) further distinguished between local link formation and distant link formation when studying new link formation with new partners. Local link formation and distant link formation were measured through the calculation of the clustering coefficient and the average distance between actors respectively.

Lin et al. (2007) proposed a method to quantify community dynamics, in which each community is represented as a vector in the interaction space and its evolution is determined by a novel interaction correlation method. Chi et al. (2007) also proposed a different approach to community detection based on both the structural and temporal analysis of the interactions between nodes. A community is defined to be a set of nodes that interact more closely with each other and this is captured by the structural analysis. However, there is a second component to communities which are the sustained interactions or interests between nodes over time. This is accounted for by considering the temporal nature of these interactions. Their method is based on factorizing the tensor matrix that captures interactions between nodes over time. Chi et al. (2008) extended the spectral clustering algorithms for evolving social network graphs and blog graphs.

Leskovec et al. (2008) proposed a model of network evolution consisting of node arrivals, edge initiation, and edge destination selection processes based on four large online social networks. They found that nodes arrive at a prespecified rate, edge initiations follow a "gap" process and edge destination selections follow triangle-closing model. Du et al. (2009) proposed the first utility-driven graph generator for weighted time-evolving networks based on the patterns that cliques follow, like the Clique-Degree Power-Law (CDPL) and Clique-Participation Law (CPL) and the observation of the weights on the edges of triangles followed power laws.

The nodes and their links in a social network change over time. Predicting changes of a social network is referred to the link prediction problem (Liben-Nowell and Kleinberg, 2007). Popescul and Ungar (2003) developed a prediction system by statistical learning that extended inductive logic programming. Their system learnt link patterns from queries of a relational database, including joins, selections and aggregations. Taskar, Abbeel and Koller (2004) applied relational Markov models to learn patterns of cliques and transitivity in web pages and hyperlinks. Both these prediction systems included node attributes, e.g., web page text, in addition to relational features. This made them more powerful than prediction systems using only topological metrics. Popescul and Ungar (2004) enhanced link prediction of author or document bipartite networks by using clustering. They clustered documents by topic and authors in order to generate new entities that were used in logistic regression of features and relations. Their system was tested on data consisting of an equal number of positive and negative cases. Zhou and Scholkopf (2004) studied classification, ranking and link prediction problems in graphs by defining discrete calculus for

graphs and then shifting classical regularization from the continuous case to graph data.

Huang, Li and Chen (2005) used link prediction to improve collaborative filtering in recommender systems. They found that the Katz measure was the most useful, followed by preferential attachment, common neighbors and measure. These path-based and neighbor-based measures outperformed simpler metrics. Farrell, Campbell and Myagmar (2005) used link prediction to design a system that recommended new academic links for researchers at a computer science conference and received feedback through a survey. They found that established researchers found little use for the system but newer researchers found it useful in recommending potential colleagues and talks of interest. Farrell et al. (2005) developed a relationship-network application and believe that social network systems will be useful to help humans cope with the huge number of professional contacts they need to maintain at present. Rattigan and Jensen (2005) presented the anomalous link discovery problem. This involves detecting which links in a network have two linked nodes that have very few common neighbors or are a large distance apart in the previous time stamp. The anomalous link discovery could be used to discover the links arising between individuals that could indicate criminal collaboration. Popescul and Ungar (2003) used the link detection problem as a test application for their work on structural logistic regression.

Coffman et al. (2004) analyzed dynamic network for the global activities of terrorists and other organized criminal groups and found that their structure and behavior differ widely from normal social networks. Fellman and Wright (2004) found that criminal networks are trade efficient for secrecy in structure and have

unusual patterns of communication. Carley (2004) proposed a model of criminal networks using dynamic techniques and created a dynamic network program, *DyNet*, where multiple agents model the social behavior of human beings, with access to resources and organizations. Holme (2003, 2004) investigated the change of the structure of social networks over time, including studies on the changing metrics of an Internet dating network and the trends of aggregate graph measures, such as the average path length and average degree.

The social network analysis and social clusters of the above work are mainly based on blogs, emails, and the World Wide Web (Wang and McCallum 2006; Kumar 2004; Kumar 2006).

2.5 Conclusion

This chapter introduced fundamental concepts, definitions and measures in social network analysis. I also reviewed and discussed a number of approaches in the areas of social group identification, network evolution and applications of social network analysis. The social groups are identified by centrality and clustering coefficient based methods based on graph theory. The difference of relationship between one member and another in the same group is not considered. In real world the relationship between one person and another is different and can dynamically change over time even though they are in the same group. In network evolution the existing approaches only consider the link formation, deletion and prediction but have not considered the quantitative changes of the relationship between one person and another. I will show the solutions for these issues in the following chapters.

CHAPTER 3

OVERVIEW OF THE PROPOSED APPROACH

Socioscope is a platform for analyzing the properties of social network structures and human social behavior, measuring interpersonal relations in groups, and discovering special events based on human telecommunication patterns.

Pentland (2006; 2007) used the mobile phones programmed, electronic badges and microphones as a Socioscope to sense and capture human behavioral data (e.g., location, proximity, body motion). These behavioral data are used to analyze the characterization of group distribution and variability, conditional probability relationships between individual behaviors and focuses on human relationship analysis based on physical distance proximity (Eagle, Pentland and Lazer, 2009). Eagle (2008) extended this approach (Pentland, 2006; 2007) to study a variety of human culture as a culture lens. My approach focused on quantifying human behaviors, interpersonal relationships, relationship changes by studying human calling patterns. Next I described the socioscope model and its components.

3.1. Socioscope Architecture

As presented in Figure 3.1, the socioscope model consists of several components including data extraction and transformation, network visualization, zoom, scale, and analysis tools which are used for analyzing network structures, discovering and quantifying social groups and events, quantifying relationships. It is extensible and new tools can be added as additional features are identified. The model is composed of three layers which were briefly described below.

Data processing: This layer consists of two components: *data extraction* and *data transformation.* In data extraction the related information is extracted from raw datasets and then transformed into required data format in data transformation for visualization and analysis.

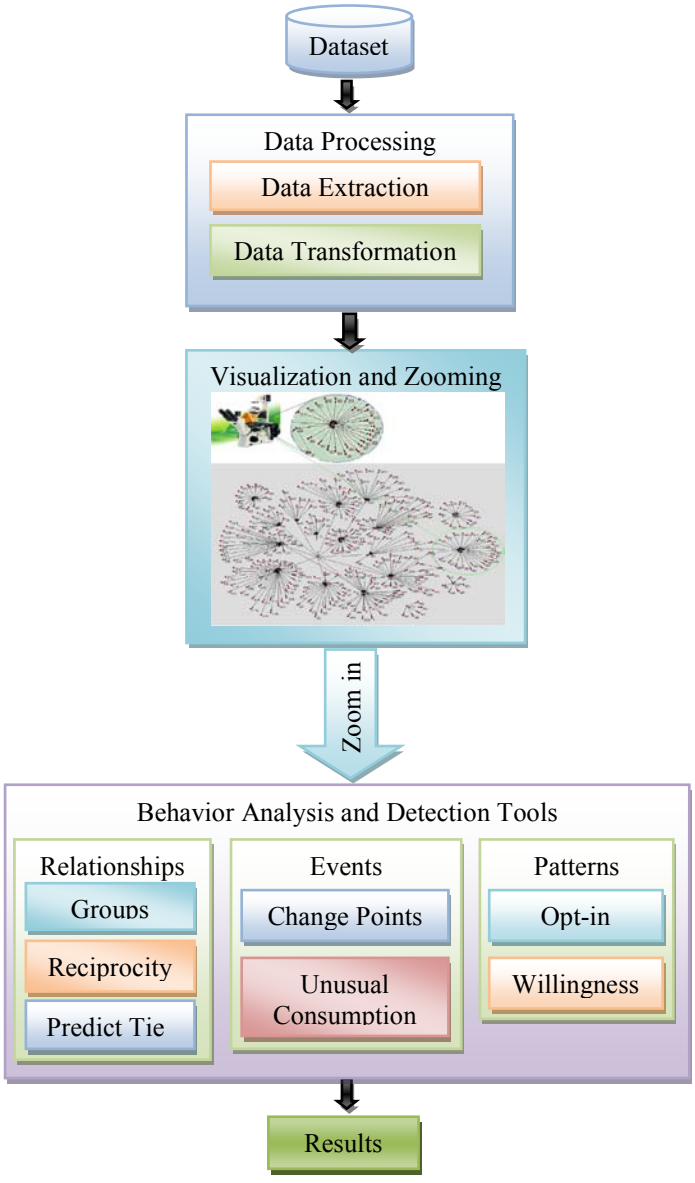


Fig.3.1 The architecture of the socioscope.

Visualization and zooming: Open source visualization tool is used for drawing the social networks. In zooming in levels multiple scales may be used to analyze

social group member behavior up to one-to-one. In zooming out levels general social network structures and properties may be analyzed.

Behavior analysis and detection tools: This layer is the core of the model and consists of several components: *quantifying social groups; reciprocity and predicting social tie strengths; detecting change points* and *unusual consumption events; detecting opt-in patterns* and *inferring willingness levels*. I described these components and solutions in detail in next chapters.

3.2. Real-life Data Sets and Parameters

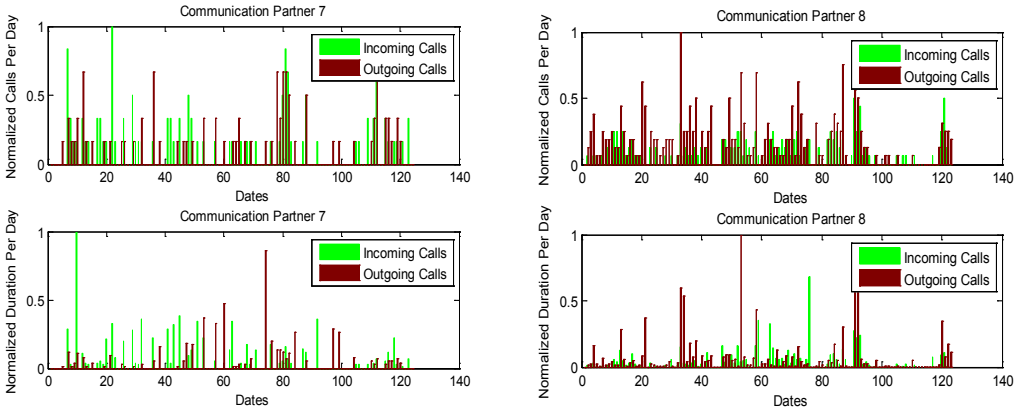
Real-life traffic profile: In this paper, actual call logs were used for analysis. These actual call logs of 81 users were collected at Massachusetts Institute of Technology (MIT) (Reality Mining 2009) by the Reality Mining Project group over a period of 8 months. Additionally, the actual call logs of 20 users were collected by the network security team at University of North Texas (UNT) over a period of 6 months.

The Reality Mining Project group collected mobile phone usage of 81 users, including user ID (unique number representing a mobile phone user), time of call, call direction (incoming and outgoing), incoming call description (missed, accepted), talk time, and tower ID (location of phone user). These 81 phone users were students, professors and staff members. The collection of the call logs was followed by a survey of feedback from participating phone users about their behavior patterns such as favorite hangout places; service providers; talk time minutes; and phone users' friends, relatives and parents.

Hand Labeling: To evaluate the accuracy of our methods we randomly chose 20 phone users from the MIT data set. The UNT data set has been used as a pattern

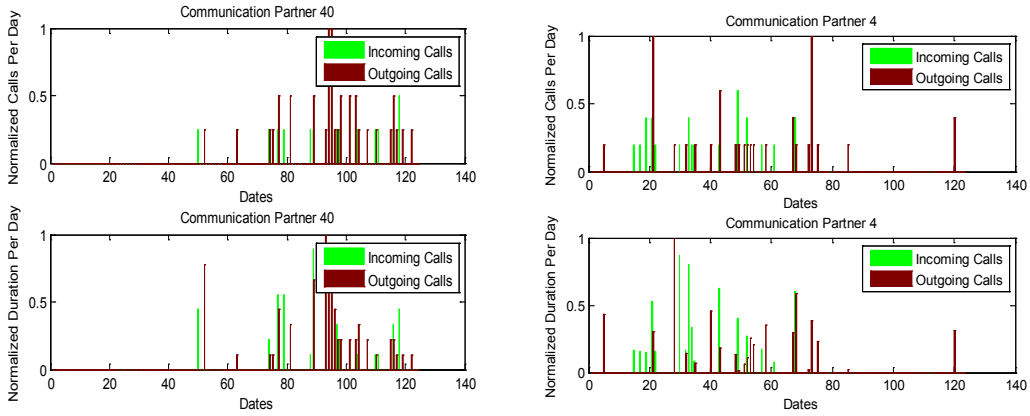
for the hand-labeling of the MIT data. That is, UNT data set contains direct feedback from users identifying their socially related groups. The identified patterns are then used to hand-label MIT data and validate our models and methods. I hand-labeled the communication members based on the number of calls, duration of calls in the period, history of call logs, location, and time of arrivals.

Figs. 3.2-3.4 show how I hand-labeled the members in MIT data, where the socially close, near and far members were obtained from user's feedback in UNT data on the left hand side figures (a) and the correspondent ones by hand-labeling the MIT data on right hand side figures (b); the *x*-axis indicates the days and the *y*-axis indicates the normalized number of calls and call duration respectively.



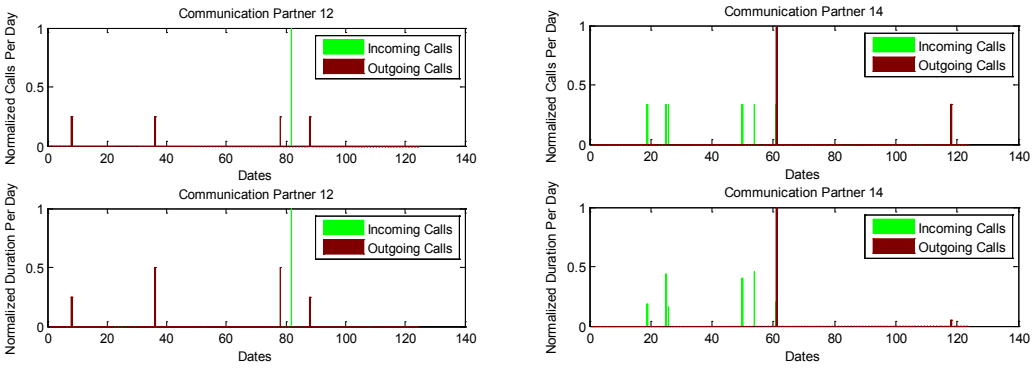
(a) (b)

Fig. 3.2 (a) Socially close member of user 12 by feedback from UNT data; (b) Socially close member of user 70 by hand-labeling from MIT data.



(a) (b)

Fig. 3.3 (a) Socially near member of user 12 by feedback from UNT data; (b) Socially near member of user 70 by hand-labeling from MIT data.



(a) (b)

Fig. 3.4 (a) Socially far member of user 12 by feedback from UNT data; (b) Socially far member of user 70 by hand-labeling from MIT data.

3.3. Conclusion

This chapter presented the socioscope architecture and the associated components for human relationship analysis based on mobile phone call records. I also described the datasets and validation method for my approach.

CHAPTER 4

QUANTIFYING SOCIAL GROUPS AND PREDICTING SOCIAL-TIE STRENGTHS

4.1 Introduction

Most social network research and social relationship analysis are based on blogs, emails and the World Wide Web (Kumar et al. 2006; Wang and McCallum 2006). Since mobile phones have become the main communication media for people in recent years, some researchers' interests in social networks concentrate on social relationship analysis based on call detail records (Nanavati et al. 2006; Onnela et al. 2007; Kurucz et al. 2007; Teng and Chou 2007; Palla et al. 2007; Hidalgo and Rodriguez-Sickert 2008; Dasgupta et al. 2008; Candia et al. 2008; Eagle et al. 2009).

Nanavati et al. (2006) and Onnela (2007) investigated the structure and tie strength of mobile telephone call graphs. Onnela et al. (2007) investigated the relationship between local network topology and the associated weights which represent the strength of social ties by the aggregate call duration and the cumulative number of calls between the individuals. They measured the number of common neighbors by the link overlaps for identifying the interconnectedness of communities. Their results provided quantitative evidence for the weak ties hypothesis. Kurucz et al. (2007) applied the spectral clustering method to telephone call graph partition. Teng and Chou (2007) discovered the communities of mobile users on call detail records using a triangle approach. Palla et al. (2007) developed the algorithm based on clique percolation to investigate the time dependence of the overlapping social groups so as to discover relationships characterizing social group evolution and capturing the

collaboration between colleagues and the calls between mobile phone users. Eagle et al. (2009) performed a new method for measuring human behavior, based on contextualized proximity and mobile phone data, to study the dyadic data using the nonparametric multiple regression quadratic assignment procedure (MRQAP), a standard technique to analyze social network data (Baker and Hubert 1981; Krackhardt 1988), to discover behavioral characteristics of friendship using factor analysis and predict satisfaction based on behavioral data. Hidalgo and Rodriguez-Sickert (2008) studied the stability of social ties by defining and measuring the persistence of the ties. They showed that the persistence of ties and perseverance of nodes depend on the degree, clustering, reciprocity and topological overlap. Dasgupta et al. (2008) proposed a spreading activation-based technique to predict potential churners by examining the current set of churners and their underlying social network. Candia et al. (2008) investigated the spatiotemporal anomalies of calls and patterns of calling activity using standard percolation theory tools. They found that interevent time of consecutive calls is heavy-tailed.

Almost all above research focused on general structures and properties for social networks. In real life we are usually interested in behaviors and quantitative relationships of some special groups of people. For example, in marketing, if someone buys something, his/her family members and friends are likely to have the same interests to buy the same or similar product and have a similar level of income although we do not know how much they earn. So we may find potential buyers through social groups. Another important application for social groups is national security. For example, if somebody is a terrorist or robber, his/her close friends or socially close communication partners are likely (not necessary) to be terrorists or

robbers. One more application is to quantify the telecommunication presence. On different days and at different times people usually would like to communicate with different groups of people. For example, we prefer to communicate with our colleagues during work time and communicate with our family members, relatives and friends during non-work time. Further, in our busy hours we only would like to have necessary communications with our socially close members such as family members, bosses. Additionally, we may be able to detect unwanted calls (e.g., spam) by social group analysis. For example, the spammers are definitely socially far from us. The system may not let the phone ring and forward the calls from socially distant groups to the voice mail box automatically.

I focused on quantifying individual social groups using a probability model, *affinity* (Fannes and Spincemaile 2003), based on mobile phone call detail records. I used affinity to measure the similarity between probability distributions, and quantified social tie's strength between actors in groups which are different from the previous work on the measurements and general structures of social networks. Since phone calls are stochastic processes, it is more suitable to use probability affinity to quantify the social relationships.

The approach proposed here for the social group identification relies first on the computation of a reciprocity index that is then used to compute the affinity between two users, and finally used to define socially related groups. A social network dynamically changes since the social relationships (social ties) change over time. The evolution of a social network mainly depends on the evolution of the social relationships. The social-tie strengths of person-to-person are different from one another even though they are in the same groups. Therefore I investigated the

evolution of person-to-person social relationships, quantified and predicted social tie strengths based on call-detail records of mobile phones. These steps were presented in the next sections.

4.2 Dyads and Reciprocity Index

In social networks, one of the important relationships between people is reciprocity. Reciprocity can be defined as the action of returning of similar acts (Katz and Powell 1955; Wasserman and Faust 1994). In this study, my interest was to investigate how people utilize technology to construct their social relationships. I focused on the measure of mediated interactions considering the media used to interact. To investigate how people interactively construct their social relationships, I focused on the reciprocity of actions that take place in a social media environment.

Reciprocity plays an important role in economic and social relations. For example, in marketing, sellers sell products to buyers. Buyers receive products and sellers earn money. Furthermore, buyers give the sellers feedback. By the buyers' feedback, the sellers improve their product quality and service. As a result, buyers receive better products and service, and sellers earn more money. Therefore, the reciprocity relation is one of the keys to business success. Similarly, we may be able to delete unwanted calls (e.g., spam) by reciprocity analysis. For example, the spammers definitely do not receive responses from us.

Gouldner (1960) proposed an index of mutuality to measure tendency toward mutuality by the probability of mutual choices between two actors. Katz and Wilson (1956) proposed an index to measure the tendency for mutuality, which compared the observed number of mutual connections to the number expected if choices were randomly made. The formulas for the mean and variance of the number of mutual

connections are given. The observed number of mutual connections is then compared to the expected number and a z-score is calculated. Schnegg (2006) found that if he includes the effect of the reciprocity and the scaling exponent, which are negatively correlated in simulations of growing network, the degree distributions are much closer to those empirically observed. Garlaschelli and Loffredo (2004) proposed a framework in which the occurrences of mutual links depended on conditional connection probabilities according to their actual degree of correlation between mutual links. Zamora-López et al. (2008) found that the 1-node and 2-node degree correlations are very important to reciprocity in real networks, and the level of correlation contributions to the reciprocity depends on the type of correlations involved. Floría et al. (2009) investigated the lattice reciprocity mechanisms and interpreted the onset of lattice reciprocity as a thermodynamic phase transition to enhance evolutionary survival of the cooperative phenomena in social networks. Hogan and Fisher (2006) found that reciprocity in email behavior is different between multi-recipient and dyadic mail.

My approach for reciprocity is different from the above work. It is observed that the structure and transactions in reciprocity are different compared to face-to-face interactions. *The existing approaches measure the tendency of mutual choices for actors (nodes) in a graph. They do not deal with frequency and duration of real time electronic communications between two actors. In real life, the frequency of communication plays an important role for the relationship between persons. To the best of our knowledge no similar work has been reported. I propose a new reciprocity index based on mobile phone call detail record*.

Most social relationship research focuses on the collection of dyads in social networks (Wasserman and Faust 1994). The new reciprocity index proposed here is different from the previous work as it also accounts for time and duration of the communication.

The dyadic relationship in a social network is the collection of dyads. A dyad is an unordered pair of nodes (actors) and arcs (ties) between the two nodes. There are *(n × (n-1))/2* dyads in a directed graph with n nodes. A dyad is mutual if both the tie from *i* to *j* and the tie from *j* to *i* are present. Each of the dyads in the network is assigned to one of three types: mutual (actor *i* has a tie to actor *j* and actor *j* has a tie to actor *i*), asymmetric (either *i* has a tie to *j* or *j* has a tie to *i*, but not both), or null (neither the *i* to *j* tie nor the *j* to *i* tie is present). These are often labeled *M*, *A*, and *N* respectively. The dyad census gives the frequencies of these types.

In (Katz and Powell 1955) the authors proposed an index of mutuality, ρ_{kp}. This index focused on the probability of a mutual choice between two actors i and j:

$$P(i \to j \,\&\, j \to i) = P(i \to j) \times P(j \to i \,|\, i \to j)$$

$P(j \to i \,|\, i \to j)$ can be considered as consisting of two parts: the $P(j \to i)$ and a fraction, denoted by ρ_{kp} of the probability $P(j \xrightarrow{not} i)$ [Katz and Powell 1955]. The ρ_{kp} is 0 if there is no tendency toward mutuality and 1 if there is a maximal tendency toward mutuality. A negative value of index indicates a tendency away from mutuality, toward asymmetry and nulls, referred to as antireciprocation. There are two kinds of ρ_{kp}, fixed choice and free choice. Fixed choice assumes that all actors make the same number of choices, and the estimate of ρ_{kp}^{fixed} is computed by (Katz and Powell, 1955)

$$\hat{\rho}_{kp}^{fixed} = \frac{2(n-1)M - nc^2}{nc(n-1-c)}$$

where n is the number of nodes, M is the observed number of mutualities, and c is the number of choices.

Free choice allows different numbers of choices and the estimate of ρ_{kp}^{free} is computed by [Katz and Powell 1955]

$$\hat{\rho}_{kp}^{free} = \frac{2(n-1)^2 M - S^2 + S_2}{S(n-1)^2 - S^2 + S_2}$$

where n is the number of nodes, M is the observed number of mutual connections, $S = \sum x_{i+}$ is the total number of choices and $S_2 = \sum x_{i+}^2$ is the sum of squares of the choices made by each actor.

In the mobile phone social networks, actor i and actor j may call each other multiple times, and the reciprocity reflects their relationship over a period of time. The above mutual index and other existing mutual indices cannot measure this kind of relationship. The existing mutual (reciprocity) indices measure the tendency of mutual choices for actors (nodes) in a graph. They do not deal with frequency of communication. I proposed a reciprocity index, $\rho_{a \leftrightarrow b}$ to measure the tendency of reciprocity for actors a and b in a group. Fig.4.1 shows the reciprocity relation between phone user 29 and his/her communication partner 349 where m, h, d and numbers inside the boxes above the arrows indicate minute, hour, day, and call duration respectively.

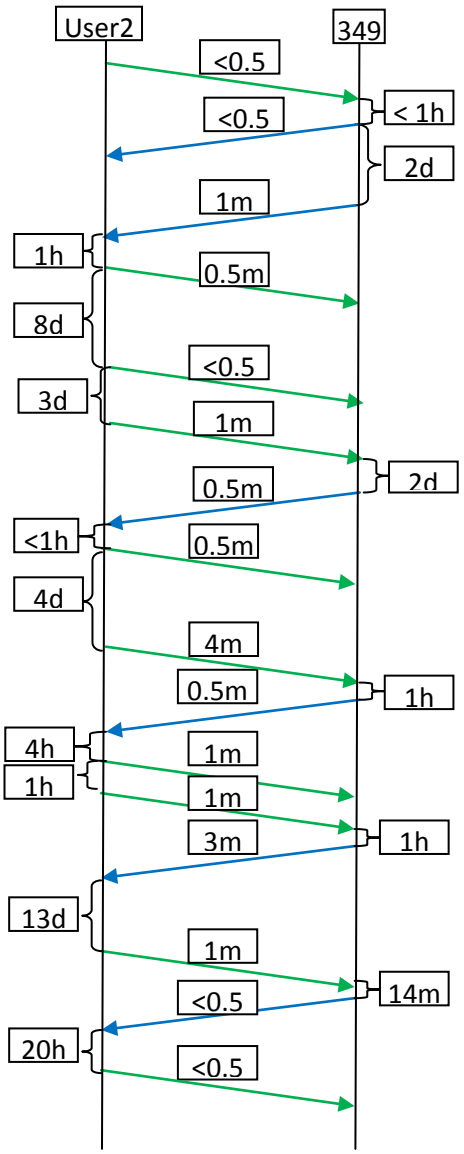


Fig. 4.1 The event flow chart for phone user 29 with his partner 349.

Bregni et al. (2006) showed that phone call arrivals follow Poisson distribution in cellular networks. Suppose that the number of phone call arrivals is a Poisson process. Then the probability of no arrivals in the interval [0, t] is given by

$$P(\tau > t) = e^{-\lambda t}$$

where λ is the arrival rate and τ is interarrival time. The occurrence of at least one arrival between 0 and t is given by

$$P(\tau \le t) = 1 - e^{-\lambda t}$$

Considering actor *a* calls actor *b* at time t_i with rate $\lambda_a t$, the probability of actor *b* calling actor *a* back (reciprocity) at a time t_j with rate $\lambda_b t$ can be computed by

$$\begin{aligned} &P(a \to b \,\&\, b \to a) = P(a \to b)P(b \to a \mid a \to b) \\ &= P(a \to b)[P(b \to a) + \rho_{a \leftrightarrow b} P(b \xrightarrow{not} a)] \\ &= (1 - e^{-\lambda_a t_i})[(1 - e^{-\lambda_b (t_j - t_i)}) + \rho_{a \leftrightarrow b} e^{-\lambda_b (t_j - t_i)}] \end{aligned}$$

The expected value, $E(R \mid \rho_{a \leftrightarrow b})$, of number of reciprocity from *b* to *a* is the total number of calls, S, from *a* to *b* times this probability, i. e.

$$E(R \mid \rho_{a \leftrightarrow b}) = S(1 - e^{-\lambda_a t_i})[(1 - e^{-\lambda_b (t_j - t_i)}) + \rho_{a \leftrightarrow b} e^{-\lambda_b (t_j - t_i)}]$$

After rearranging the terms, we have

$$\rho_{a \leftrightarrow b} = [R - S(1 - e^{-\lambda_a t_i})(1 - e^{-\lambda_b (t_j - t_i)})] / S(1 - e^{-\lambda_a t_i}) e^{-\lambda_b (t_j - t_i)} \qquad (4.1)$$

where R is observed number of reciprocity.

The $\rho_{a \leftrightarrow b}$ is 0 if there is no tendency toward reciprocity and 1 if there is a maximal tendency toward reciprocity.

I computed the reciprocity indices by Eq. (4.1) for the call log data. In this paper I defined that the reciprocity time interval $t_j - t_i$ is 24 hours, i.e., the returned calls or messages within a 24-hour period were used to compute the reciprocity index. This was only an example to choose $t_j - t_i$. This parameter can be adjusted to any reasonable length of time.

4.3 Social Group Identification

Groups correspond to clusters of data. Cluster analysis concerns a set of multivariate methods for grouping data variables into clusters of similar elements. In this paper I used probabilistic models for the classification of variables by the *affinity* (Fannes and Spincemaile 2003).

The observed data can be represented in a bi-dimensional matrix, where rows describe data units and columns describe categorical variables. Empirical clustering models are usually used to analyze such data. In the first step one has to choose an appropriate proximity (similarity or dissimilarity) coefficient to measure the relationship between pairs of elements within the data set to classify. In the second step one has to define an aggregation criterion for merging similar clusters of elements, and in the third step one has to use some criteria to assess the validity of the clustering results. In this paper I applied more appropriate probabilistic model. In the first step I applied probabilistic similarity coefficient, i.e. *affinity*, that is measured in a probability scale, instead of simple/basic similarity coefficients. In the second step I defined an aggregation criterion for merging similar clusters of elements. In the third step I used some ways to assess the validity of the clustering results. I used internal validation, i.e. similarity coefficients, to compare a classification with the original data sets.

Affinity measures the similarity between probability distributions. Since my problem belongs to discrete events, I only considered finite event spaces. Let

$$S_N = \{P = (p_1, p_2, ...p_N) \mid p_i \geq 0, \sum_{i=1}^{N} p_i = 1\}$$

be the set of all complete finite discrete probability distributions and $P,Q \in S_N$. The Hellinger distance between P and Q is defined as (Fannes and Spincemaile 2003)

$$d_H^2(P,Q) = \frac{1}{2}\sum_{i=1}^{N}(\sqrt{p_i} - \sqrt{q_i})^2 \tag{4.2}$$

$d_H^2(P,Q) \in [0,1]$, $d_H^2(P,Q) = 0$ if P = Q and $d_H^2(P,Q) = 1$ if P and Q are disjoint.

The affinity between probability measures P and Q is defined as (Fannes and Spincemaile 2003)

$$A(P,Q) = 1 - d_H^2(P,Q) = \sum_{i=1}^{N}\sqrt{p_i q_i} \tag{4.3}$$

$A(P,Q) \in [0,1]$, $A(P,Q) = 1$ if P = Q and $A(P,Q) = 0$ if P and Q are disjoint.

Proof:

$$\begin{aligned} A(P,Q) &= 1 - d_H^2(P,Q) = 1 - \frac{1}{2}\sum_{i=1}^{N}(\sqrt{p_i} - \sqrt{q_i})^2 \\ &= 1 - \frac{1}{2}\sum_{i=1}^{N}(p_i - 2\sqrt{p_i}\sqrt{q_i} + q_i) \\ &= 1 - \frac{1}{2}(\sum_{i=1}^{N} p_i - 2\sum_{i=1}^{N}\sqrt{p_i q_i} + \sum_{i=1}^{N} q_i) \\ &= \sum_{i=1}^{N}\sqrt{p_i q_i} \end{aligned}$$

For finite and discrete data, let *M(X,Y)* be a $L \times N$ matrix, where *X* represents the set of data units and *Y* is a set of *N* categorical variables. In this paper Y_j (*j=1,...,N*) is a vector of frequencies. Thus Y_j may be the represented by the *L* coordinates n_{ij} (*i=1, 2, ..., L*) which is a frequency. We will refer to the j-th column profile as the corresponding conditional vector with $n_{ij} \Big/ \sum_{i=1}^{L} n_{ij}$. This profile vector may be a discrete conditional probability distribution law. It is often a profile or probability vector of the population, where the set *X* of *L* data units represent a partition of some

random sample of subjects in L classes. In this paper $p_i = n_{ij} \Big/ \sum_{i=1}^{L} n_{ij}$. The column profiles have a major role since the similarity between pairs of variables will be measured using an appropriate function, the affinity in this paper, of their profiles.

In our life we have relationships with a small group of individuals in our social network such as family members, relatives, friends, neighbors and colleagues. We divide the time of a day into working time (8am-5pm) and nonworking time (5:01pm-7:59am). Note that since the data was collected from students, professors and staff members in the university, the work time may be different from the regular work time (8am-5pm) which was found by our previous work in (Zhang and Dantu 2008), but I still used the regular work time for generalization. Further, in the two time periods people divide their social network members into three categories: socially close members, socially near members and socially far members.

- *Socially close members:* These are the people with whom we maintain the strongest social relationship. As reflected by phone calls, we receive more calls from them and we tend to talk to them for longer period of time. Family members, close friends and colleagues in the same team belong to this category.
- *Socially near members:* These relationships are not as strong as those of family members, close friends and colleagues in the same team. Sometimes, not always, we connect to each other and talk for considerably longer periods. We mostly observe intermittent frequency of calls from these people. Distant relatives, general friends, colleagues in a different team and neighbors are in this category.

- *Socially far members:* These people have weaker relationships with each other in social life. They call each other with lower frequency. We seldom receive calls from them and talk to them for a short period of time.

In this paper, I used three attributes: incoming (*in*), outgoing (*out*), and reciprocity (*reci*) of calls and messages.

Let m_i, n_i be the number of calls, where $i \in \{in, out, reci\}$. $P = (p_{in}, p_{out}, p_{reci})$ is a vector of normalized frequencies over the training period and $Q = (q_{in}, q_{out}, q_{reci})$ is a vector of normalized frequencies of the same attributes observed over the testing period. Then

$p_i = m_i / \sum_i m_i$ where $i \in \{in, out, reci\}$ and

$q_i = n_i / \sum_i n_i$ where $i \in \{in, out, reci\}$.

The reciprocity part is computed by Eq. (4.1).

The affinity between P and Q is computed as follows:

$$A(P,Q) = \sum_i \sqrt{p_i q_i} \text{ where } i \in \{in, out, reci\} \quad (4.4)$$

We use the actual call log data to compute the affinity values given by Eq. (4.4). We define:

- Socially close members if $0.9 < A(P, Q) \leq 1$
- Socially near members if $0.3 < A(P, Q) \leq 0.9$
- Socially far members if $0 \leq A(P, Q) \leq 0.3$

4.4 Social Tie Strength Prediction

Most of prediction applications in social networks focused on link prediction (Liben-Nowell and J. Kleinberg, 2007). Gilbert and Karahalios (2009) proposed a predictive model that maps online social network data to tie strength which is a linear

combination of the predictive variables. This model classifies friends as strong and weak ties. Kahanda and Neville (2009) used models including logistic regression, bagged decision trees, and naive Bayesian classifiers to predict strong ties in online social network based on transactional information such as communication, file transfer, email.

This study differs from previous work on measurements which focused on general structures of social networks. The social-tie strengths in existing work were not considered to be functions of time. In real world social relationship between two people changes over time.

Time series prediction plays an important role in various applications. Several schemes have been widely deployed for predicting weather, environment, economics, stock, market, earthquakes, flooding, network traffic and call center traffic (Gooijer and Hyndman, 2006). Companies use predictions of demands for making investments and efficient resource allocation. The call centers predict workload so that they can get the right number of staff in place to handle it. Network traffic prediction is used to access future network capacity requirements and to plan network development for optimum use network resources and improve quality of services.

In this study the affinity model was used to map the call-log data to social tie strengths in a time series and then the seasonal auto regressive integrated moving average (SARIMA) model was applied to predict the social tie strength of the next period.

4.4.1 SARIMA Prediction Model

The Box-Jenkins methodology is a widely used technique for determining the most appropriate ARMA, ARIMA or SARIMA model for a given time series (Box

and Jenkins, 1994).

Box-Jenkins prediction models are based on statistical concepts and principles and are able to model a wide spectrum of time series behavior. It has a large class of models to choose from and a systematic approach for identifying the correct model form.

Seasonal Auto Regressive Integrated Moving Average (SARIMA) model integrates Seasonal (periodic), autoregressive (AR), integrated (I), and moving average (MA) into a general comprehensive time series model (Box and Jenkins, 1994).

Let $\{X_t\}$ be a stationary time series.

An $AR(p)$ represents that each observation is a function of the previous p observations, which defined as

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_p X_{t-p} + e_t$$

where X_t is an observation, ϕ_i ($i=1, \ldots, p$)are coefficients to be estimated and $e_t \sim (0, \sigma^2)$ (white noise).

A $MA(q)$ describes that each observation is a function of the previous q errors, which defined as

$$X_t = e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \ldots - \theta_q e_{t-q}$$

where θ_i ($i=1, \ldots, q$)are coefficients to be estimated and $e_t \sim (0, \sigma^2)$.

$AR(p)$ and $MA(q)$ can be combined in an $ARMA(p,q)$ model

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_p X_{t-p} + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \ldots - \theta_q e_{t-q}$$

Considering the backward linear operator B defined by $BX_t = X_{t-1}$, $B^i X_t = X_{t-i}$ for any integer i.

By using the backshift linear operator, an $AR(p)$ can be written as

$$X_t - \phi_1 X_{t-1} - \phi_2 X_{t-2} - \ldots - \phi_p X_{t-p} = \phi_p(B) X_t = e_t$$

where $\phi_p(B)$ is defined by

$$\phi_p(B) = 1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p$$

Similarly, $MA(q)$ can be written as

$$X_t = e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \ldots - \theta_q e_{t-q} = \theta_q(B) e_t$$

where $\theta(B)$ is defined by

$$\theta_q(B) = 1 - \theta_1 B - \theta_2 B^2 - \ldots - \theta_q B^q$$

An $ARMA(p,q)$ model can be written in backshift linear operator form

$$(1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p) X_t = (1 - \theta_1 B - \theta_2 B^2 - \ldots - \theta_q B^q) e_t$$

That is,

$$\phi_p(B) X_t = \theta_q(B) e_t \qquad (4.5)$$

If $\{Y_t\}$ is an nonstationary time series, $ARMA(p,q)$ can be extended to $ARIMA(p,d,q)$.

$I(d)$ component removes trend by conducting d differencing operations between consecutive observations making time series stationary.

Considering the difference linear operator Δ defined by

$$\Delta Y_t = Y_t - Y_{t-1} = Y_t - BY_t = (1 - B) Y_t$$

The stationary time series $\{X_t\}$ can be obtained by the d^{th} difference Δ^d of nonstationary time series $\{Y_t\}$

$$X_t = \Delta^d Y_t = (1-B)^d Y_t \qquad (4.6)$$

By substituting X_t in (1) with (2), we have $ARIMA(p,d,q)$ model

$$\phi_p(B)\Delta^d Y_t = \theta_q(B)e_t \qquad (4.7)$$

or

$$\phi_p(B)(1-B)^d Y_t = \theta_q(B)e_t \qquad (4.8)$$

Applying a similar idea to seasonal time series, I used seasonal difference and backshift operators.

Let $\{Z_t\}$ be a seasonal time series.

The seasonal difference linear operator Δ_s defined by

$$\Delta_s Z_t = Z_t - Z_{t-s} = Z_t - B^s Z_t = (1-B^s)Z_t$$

where *s* is the length of the seasonal variation (period).

The D^{th} difference Δ_s^D of seasonal time series $\{Z_t\}$ is $\Delta_s^D Z_t = (1-B^s)^D Z_t$

A seasonal $AR(P)$ operator of order P is defined as

$$\Phi_P(B^s) = 1 - \Phi_1 B^s - \Phi_2 B^{2s} - \ldots - \Phi_P B^{Ps}$$

A seasonal $MA(Q)$ operator of order Q is defined as

$$\Theta_Q(B^s) = 1 - \Theta_1 B^s - \Theta_2 B^{2s} - \ldots - \Theta_Q B^{Qs}$$

The seasonal time series can be modeled as

$$\Phi_P(B^s)\Delta_s^D Z_t = \Theta_Q(B^s)Y_t \qquad (4.9)$$

or

$$\Phi_P(B^s)(1-B^s)^D Z_t = \Theta_Q(B^s)Y_t \qquad (10)$$

Substituting (4.7) and (4.8) into (4.9) and (4.10) respectively, yield the seasonal $ARIMA(p,d,q)$ model with period s

$$\phi_p(B)\Phi_P(B^s)\Delta^d\Delta_s^D Z_t = \theta_q(B)\Theta_Q(B^s)e_t \qquad (4.11)$$

or

$$\phi_p(B)\Phi_P(B^s)(1-B)^d(1-B^s)^D Z_t = \theta_q(B)\Theta_Q(B^s)e_t \quad (4.12)$$

which are denoted by $SARIMA(p,d,q)\times(P,D,Q)_s$.

4.4.2 Model Creation and Prediction Procedure

The Box-Jenkins method uses an iterative approach of identifying a possible model from a general class model. The chosen model is then checked against the historical data to see whether it accurately describes the series. The model fits well if the residuals are generally small. It comprises the following steps.

Step 1: Model Identification

The first step in model identification is to determine whether the series is seasonal (periodic) and stationary and whether the time series requires differencing to remove seasonality and trend. The initial parameters of an ARIMA model are based on an examination of a plot of the time series. Stationary series appears to vary about a fixed level. It is useful to look at a plot of the series along with the sample autocorrelation function (ACF). The autocorrelation function (ACF) and partial autocorrelation function (PACF) are used to identify the best model. If the ACF of the dependant variable approaches zero as the number of lags increases, the series is stationary. A nonstationary time series show little tendency for the ACF's to decrease in size as the number of lags increase. A nonstationary time series is indicated if the

series appears to grow or decline over time and the sample autocorrelations fail to die out rapidly.

If the series is not stationary, it can often be converted to a stationary series by *d* differencing. If differencing, the original series is replaced by a series of differences. An $ARIMA(p,d,q)$ model is then specified for the differenced series. The $ARIMA(p,d,q)$ models reduce to the $ARIMA(p,q)$ models.

Once a stationary series has been obtained, the form of the model to be used must be identified. This involves selecting the most appropriate lags (values for *p* and *q*) for the *AR* and *MA* parts. These are chosen by finding the lowest *p* and *q* for each residual of the estimated equation.

Step 2: Model Estimation

Once a tentative model has been selected, the parameters for that model must be estimated. The parameters in $ARIMA(p,d,q)$ models are estimated by minimizing the sum of squares of the fitting errors. This usually involves the use of a least squares estimation process. The residual mean square error is useful for assessing fit and comparing different models. This is done by

- Choose initial *p* and *q*
- Estimate *AR(p)* and *MA(q)* equations for that *p* and *q* and then chosen above, and
- Test the residuals of the estimated equation to see if they are free of autocorrelation. If they are not, either *p* or *q* is increased by one, and the process is begun again.

After initial *p* and *q* are chosen, the residuals from this estimate are analyzed with ACF and PACF. The PACF for the k^{th} lag is the correlation coefficient between e_t and e_{t-k}. In particular, the last lag before the PACF moves toward zero with an exponential decay is typically a good value for *p*, and the last lag before the ACF moves toward zero with an exponential decay is typically a good value for *q*.

After the initial *p, d* and *q* are estimated, the *ACF* and PACF of the residuals are calculated and inspected. If these ACF and PACF are all significantly different from zero, the equation can be considered as the final one since the residuals are free from autoregressive and moving average components. If one or more of the ACF or PACF are significantly different from zero, increase *p* if PACF is significant or *q* if ACF is significant by one and re-estimate the model. This process will be performed until the residuals have no autoregressive and moving average components.

Step 3: Model Validation

As a final check, it is suggested to compare the variance of $ARIMA(p,d,q)$ with those of $ARIMA(p+1,d,q)$ and $ARIMA(p,d,q+1)$. If $ARIMA(p,d,q)$ has the lowest variance, it should be considered the final equation.

Step 4: Prediction

Prediction can involve either in-sample or out-of-sample prediction, for the out-of-sample predictions, the data used for the prediction is not included in the estimation of the model used for the prediction. When assessing how well a model predicts, we need to compare it to the actual data, this then produces a prediction error (difference between prediction and actual values) for each individual observation used for the prediction. Then the accuracy of the prediction needs to be measured.

4.5 Experimental Results and Discussions

4.5.1 Quantifying Social Groups and Reciprocity index

In this section I first presented a series of individual results to show the behavior of the models and then I presented summarized results at the end.

Fig. 4.2 shows the call network of subset of call detail records in one month in which there are 326 vertices labeled by phone number ids which denote the communication members and the corresponding arcs representing the incoming or outgoing calls by the arrows. There are about 3200 communication members in the four month call detail records. Since the space is limited, I only used the call network of subset of call detail records to show the relationships among the communication members. The phone number id of user 29 is 264 and the part of his communication members is shown in Fig. 4.2.

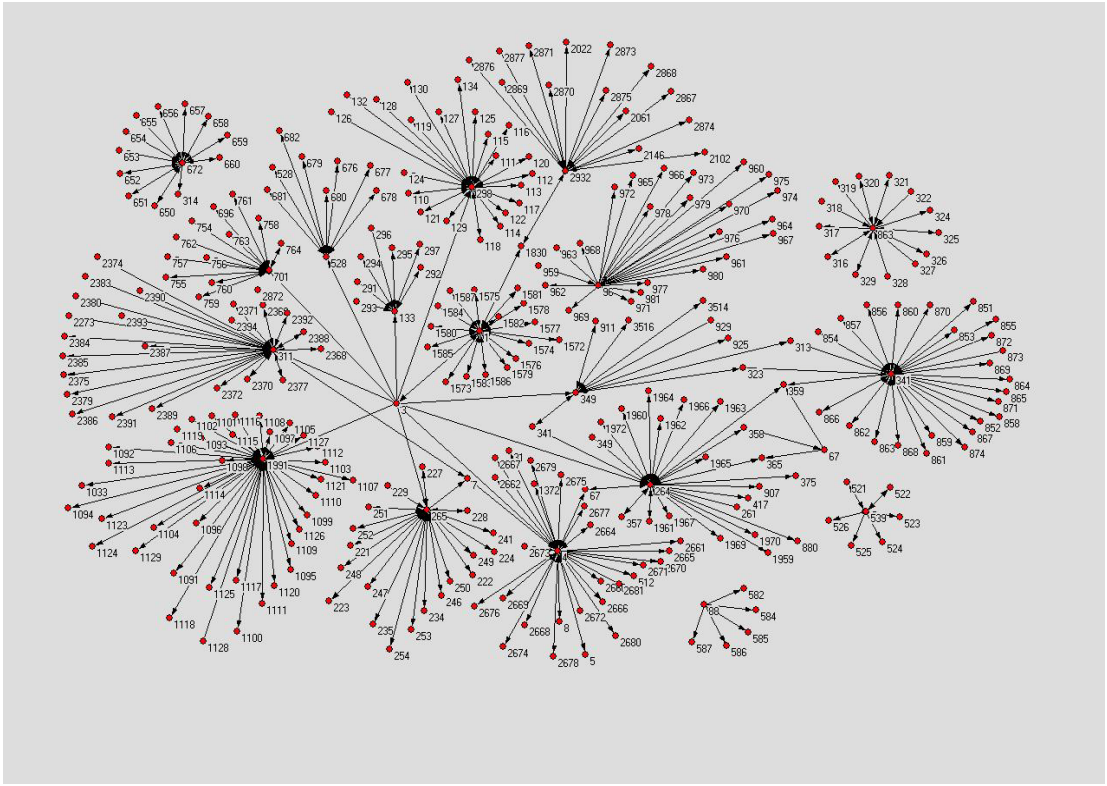


Fig. 4.2 The call network of the subset of the call detail records.

In Fig. 4.3, the x-axis indicates the phone numbers that are used to communicate with user 29 over four months, and the y-axis indicates the affinity values based on both number of calls and call duration respectively. From the Fig. 4.3 user 29 has seven socially close members, eight socially near members and twenty four socially far members in this four-month period. I divided the social group members into work time members and non-work time members. In general, during work time people prefer to talk to colleagues, bosses, secretaries, clients and customers, occasionally speak to family members and friends when needed, and during non-work time people usually talk to family members and friends, and occasionally speak to colleagues, clients and customers when needed. Note that some people may be both work time colleagues and non-work time friends. Thus the set of work time members and the set of non-work time members may overlap. User 29, who was a student, had one socially close member, two socially near members and one socially far member in work time and four socially close members, eleven socially near members and twenty-three socially far members in non-work time. Note that in work time user 29 may also use the office public phone to speak to his colleagues. In this paper I only used the cell phone call logs to classify the social groups. Since he was a student, he probably had no clients and customers, and he had only one socially close member, two socially near members and one socially far member in work time.

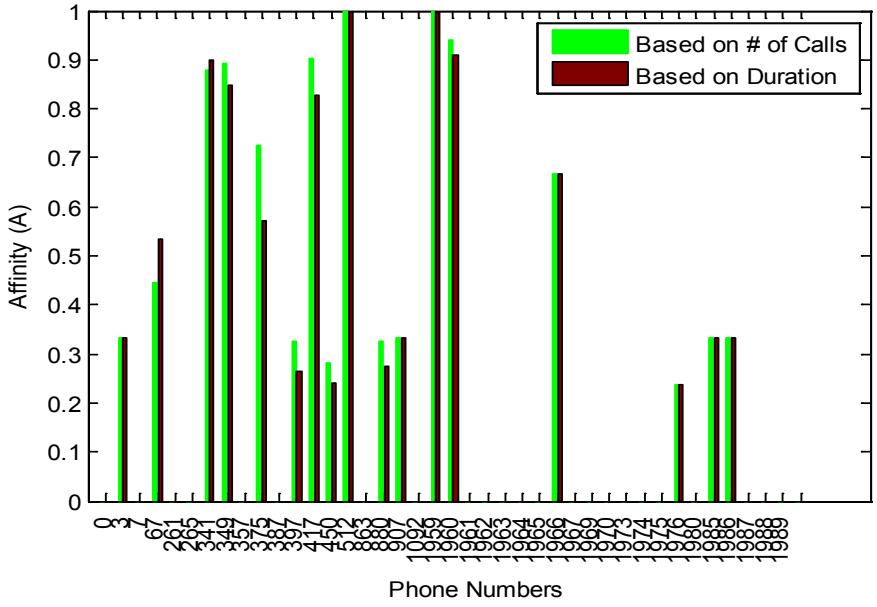


Fig. 4.3 The affinity values for phone user 29.

Fig. 4.4 shows the reciprocity index results of user 29's communication partners, where the x-axis indicates the phone numbers and the y-axis indicates the reciprocity index values. User 29 has 39 communication partners. For example, the reciprocity index is 0.72 for the communication partner 375 and 0 for the communication partner 7.

Fig. 4.5 shows the reciprocity index results of user 15 with his 29 communication partners. In most cases in the experiments, the higher reciprocity index values reflect a closer relationship between members.

To find the relationships between the reciprocity index and different time intervals of reciprocity, I calculated the reciprocity index for different time intervals $t_j - t_i$ which equals to 1, 2, …, 24 hours respectively.

Fig. 4.6 shows the reciprocity index for user 39 with his communication partner 316, where the x-axis indicates the time intervals in hours and the y-axis

indicates the reciprocity index values. Fig. 4.6 shows decreasing trend of the reciprocity index values when the time intervals increase.

Fig. 4.7 shows the probability of the reciprocity time, where the x-axis indicates the time intervals in hours, the y-axis indicates the probability and the curves are the fitted functions for user 39 with his partner 316 who is a frequent communication partner. Fig. 4.7 shows that the reciprocity time has exponential distribution and most reciprocity time is within 1 hour. I found that the reciprocity distributions follow exponential trends for frequent communication partners. I also found that the case in which the reciprocity time are greater than 10 hours mostly happened when actor *a* called actor *b* at about evening sleeping time and actor *b* called actor *a* back on the next day.

By distribution fitting I have probability density function

$$f(t) = 2.8e^{-1.5t}$$

for user 39 with partner 316.

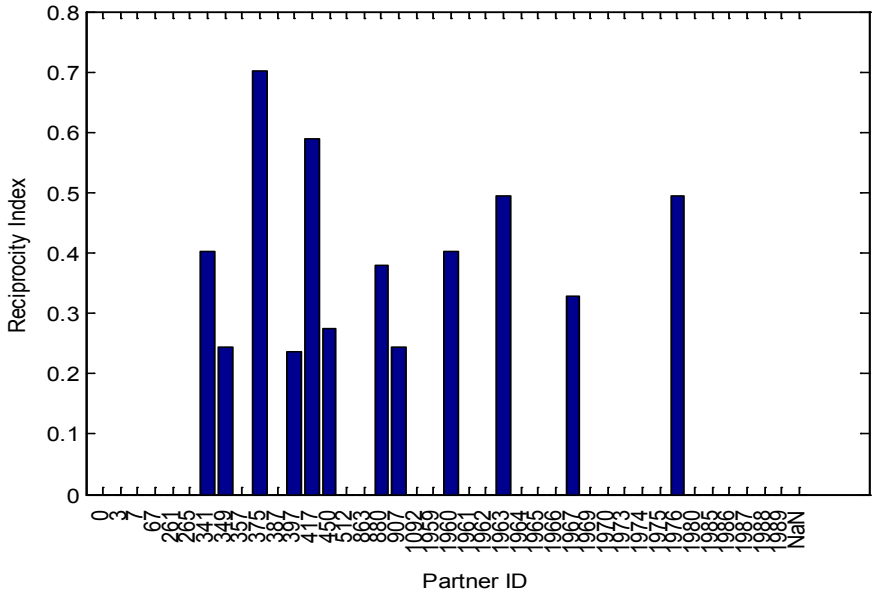


Fig. 4.4 The reciprocity indices for phone user 29.

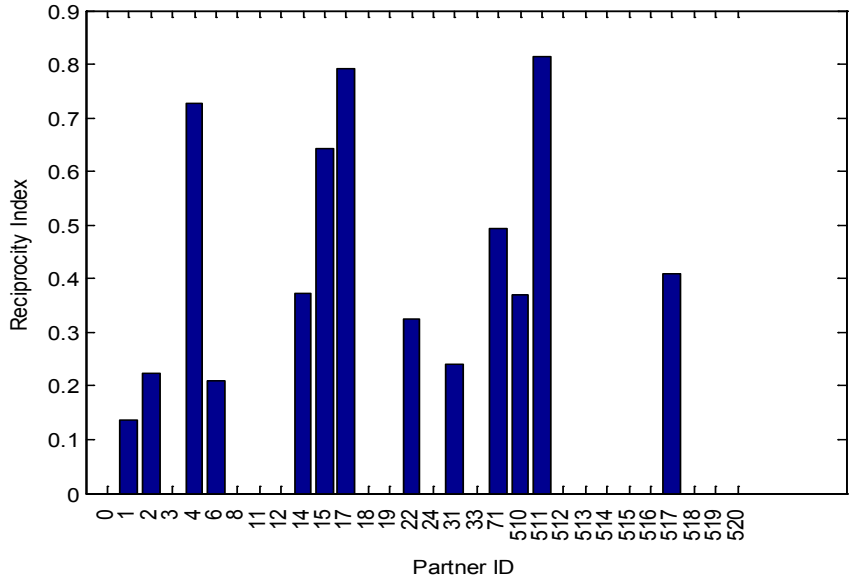


Fig. 4.5 The reciprocity indices for phone user 15.

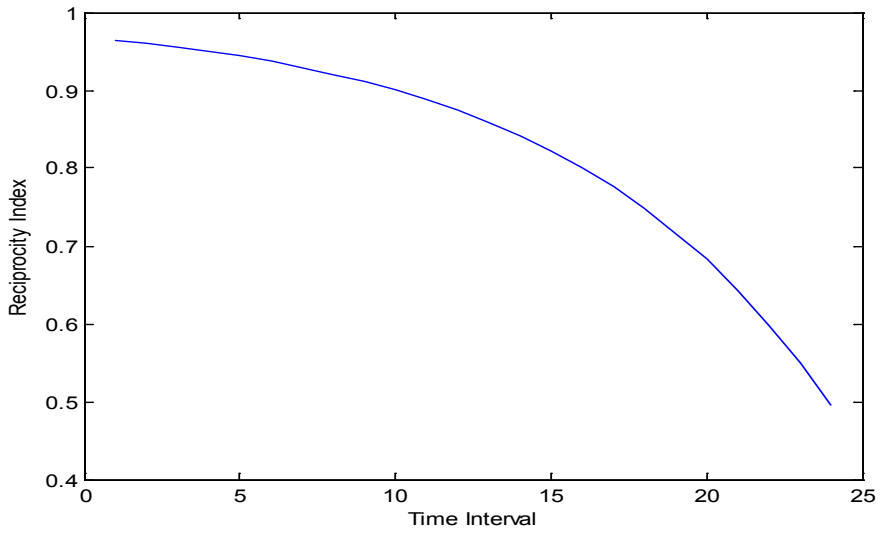


Fig. 4.6 The reciprocity index values for phone user 39 with his partner 316.

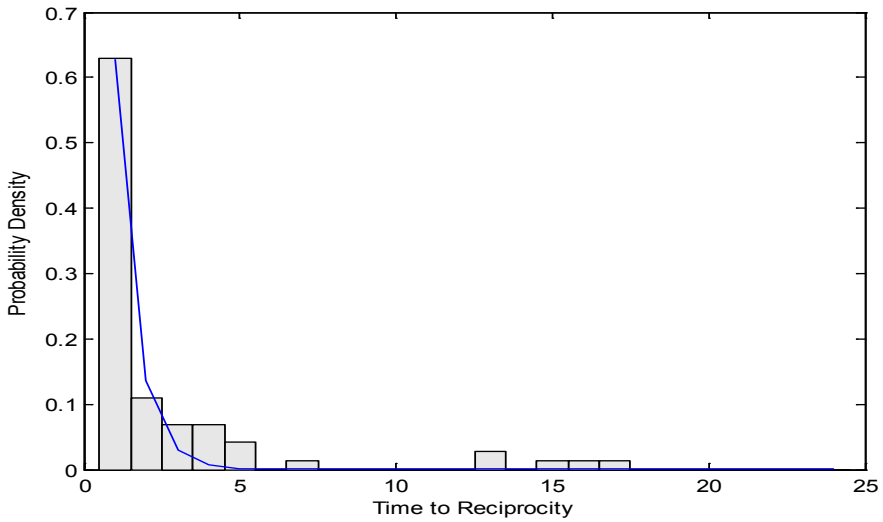


Fig. 4.7 The probabilities of reciprocity for phone user 39 with his partner 316.

Table 4.1 describes the experimental results for 20 phone users and their social groups. It is found that the failed or unsure cases only happened when the number of calls is very few and only incoming or outgoing and very consistent in each month period of time during this four month period. However, these kinds of cases seldom happened in the experiments. The affinity model achieves performance with high accuracy of 94%.

Table 4.1 Social Groups for Phone Users

User ID	Total # of members	Close	Near	Far	Hit	Fail	Unsure
user29(student)	39	4	11	24	38	0	1
user3(student)	61	5	9	47	57	1	3
user14(student)	35	4	6	25	33	1	1
user15(student)	29	5	9	15	25	2	2
user16(student)	41	4	9	28	39	1	1
user41(professor)	39	4	10	25	37	0	2
user21(student)	20	5	2	13	18	1	1
user22(student)	46	5	11	30	43	2	1
user74(student)	13	2	4	7	12	0	1

user88(staff)	66	5	9	42	63	0	3
user33(staff)	31	4	4	23	31	0	0
user35(student)	72	3	19	50	67	2	3
user38(student)	63	5	27	31	59	3	1
user39(student)	64	2	11	51	62	1	1
user70(student)	65	5	23	39	61	2	2
user49(student)	18	5	3	10	16	1	1
user50(student)	63	6	14	43	61	0	2
user57(student)	43	2	13	28		0	1
user83(student)	42	4	8	30	39	1	2
user95(professor)	8	1	4	3	8	0	0

4.5.2 Predicting Social-Tie Strengths

First the affinity values are computed and mapped in a time series in which the time unit is chosen biweekly.

The social network members are classified into socially close, near and far members. The social group results are validated by combining user's feedbacks and hand-labeling. It is found that most of errors occurred in social far members since the number of calls was few.

Since there is only a few number of calls from socially far members, the social-tie strengths are only predicted for socially close and near members.

Fig. 4.8 shows the ACF and PACF values for user 60 with his partner 2538. Fig. 4.8 shows that this time series is nonseasonal and stationary.

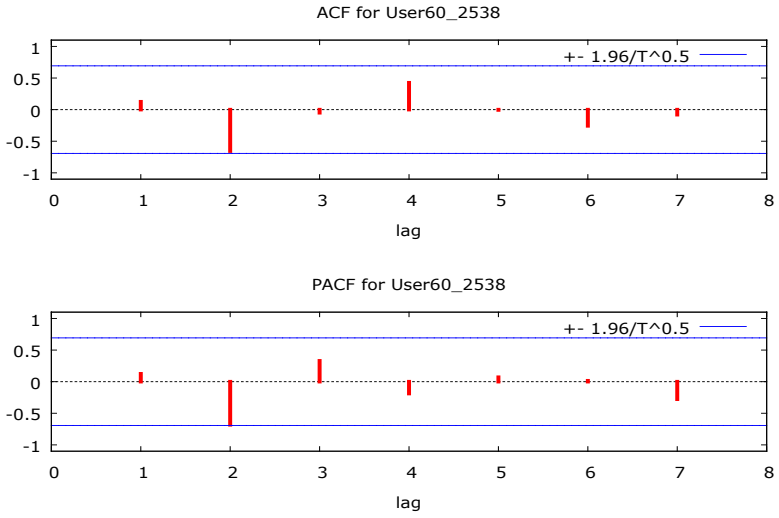


Fig. 4.8 The ACF and PACF for user 60 with his partner 2538.

From the model creation and prediction procedure described above, I obtained the model in the form $ARIMA(2,0,3)$ and

$$X_t = 0.9173 - 0.0580X_{t-1} - 0.9823X_{t-2} + 0.8866e_{t-1} + 0.8935e_{t-2} - 0.1070e_{t-3} + e_t$$

for user 60 with his partner 2538.

Fig. 4.9 shows both observed and predicted values of affinity including confidential intervals for the last 2 predicted values, where the x-axis indicates the time biweekly and the y-axis indicates the affinity values for user 60 with his partner 2538. Fig. 4.10 shows residuals of the affinity values corresponding to Fig.4.9. The square root of mean square error is 0.013.

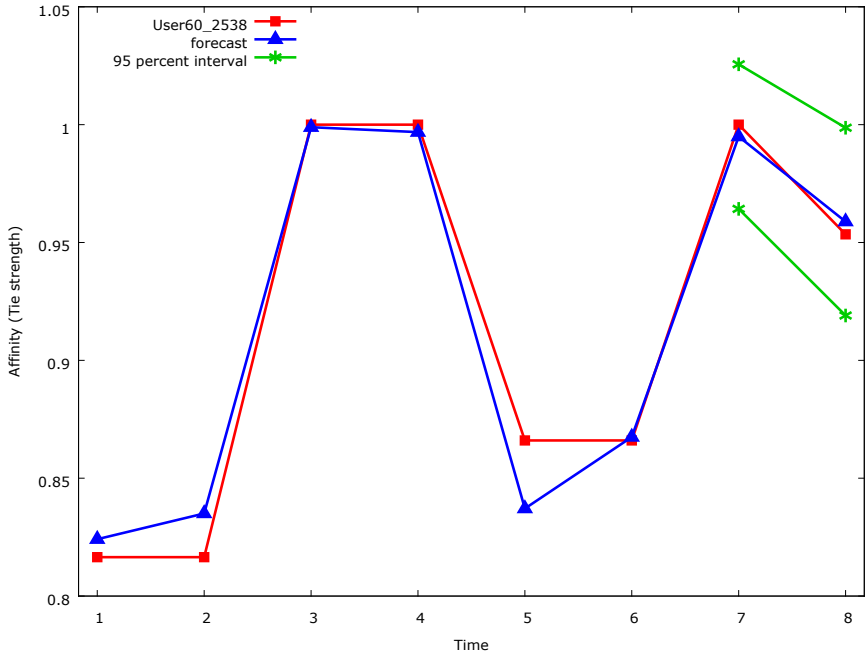


Fig. 4.9 The affinity predicted and observed values for user 60 with his partner 2538.

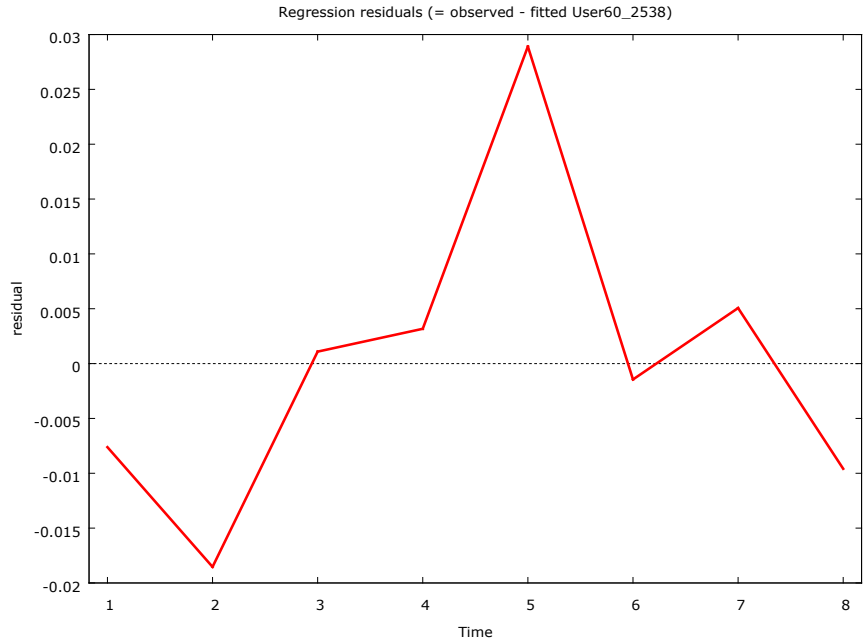


Fig. 4.10 Residuals of the affinity values for user 60 with his partner 2538.

Fig. 4.11 shows the *ACF* and *PACF* values for user 60 with his partner 2538.

Fig. 4.11 shows that this time series is nonseasonal and stationary.

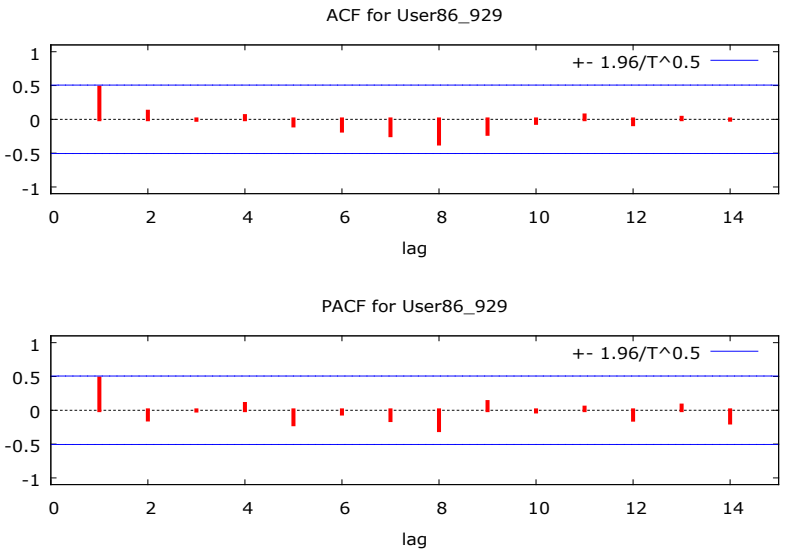


Fig. 4.11 The *ACF* and *PACF* for user 86 with his partner 929.

Fig. 4.12 shows both observed and predicted values of affinity including confidential intervals for the last 2 predicted values, where the x-axis indicates the time biweekly and the y-axis indicates the affinity values for user 86 with his partner 929.

Fig. 4.13 shows residuals of the affinity values correspondent to Fig. 4.12. The square root of mean square error for prediction is 0.076.

For user 86 with his partner 929, the model of affinity prediction is in the form $ARIMA(4,0,4)$ and

$$X_t = 0.8272 + 0.3103X_{t-1} - 0.1687X_{t-2} - 0.4026X_{t-3} + 0.4228X_{t-4}$$
$$- 0.4895e_{t-1} - 0.2072e_{t-2} - 0.4899e_{t-3} - 0.9996e_{t-4} + e_t$$

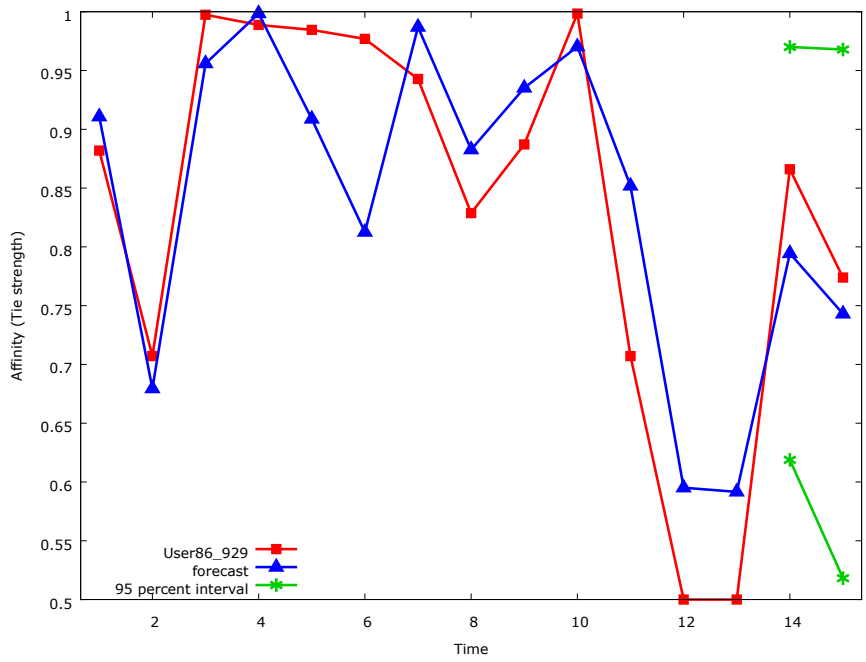


Fig. 12 The predicted and observed affinity values for phone user 86 with partner 929.

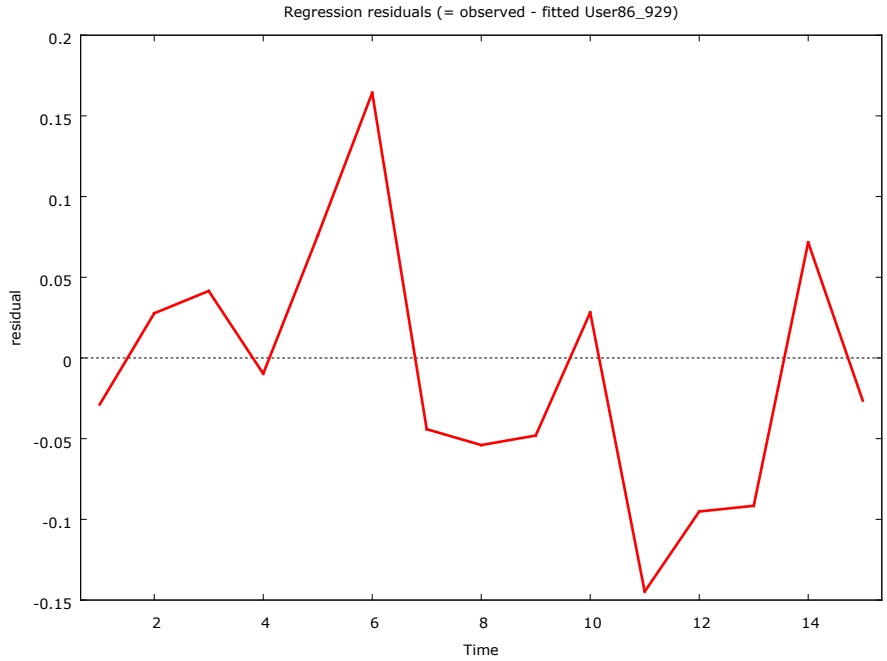


Fig. 13 Residuals of the affinity values for user 86 with his partner 929.

Table 4.2 describes the prediction errors for randomly selected phone users. The prediction achieved performance with accuracy of average 95.2%.

TABLE 4.2
Prediction Errors

User ID	Mean Error	Mean Squared Error	Root Mean Squared Error
User60-2538	0.0001	0.0001	0.0130
User60-3	0.0079	0.0001	0.0100
User60-2518	0.0245	0.0006	0.0248
User60-2519	-0.0124	0.0002	0.0144
User60-2524	-0.0011	0.00004	0.0070
User86-929	-0.0088	0.0058	0.0767
User86-341	0.00002	0.00015	0.0125
User86-911	0.0093	0.00009	0.0098
User86-3514	-0.0009	0.0000009	0.0009
User70-1661	0.0000007	0.00002	0.0052
User70-1662	0.00006	0.000002	0.0016
User70-1664	0.0599	0.0182	0.1352
User70-1667	0.0082	0.0058	0.0767
User70-1680	0.1145	0.0152	0.1233
User70-1788	0.0759	0.0113	0.1063
User50-3268	0.0106	0.0014	0.0386
User74-2906	0.0040	0.0355	0.1884
User39-316	0.0014	0.000002	0.0015
User49-2003	0.0027	0.000007	0.0027
User49-2005	-0.028	0.0148	0.1216

4.6 Conclusion

In this chapter I proposed a new *reciprocity index* for measuring the levels of reciprocity between users and their communication partners. I also proposed *affinity* model in which the new *reciprocity index* was integrated for quantifying social groups, mapping call-log data into time series and applying *SARIMA* model for predicting social tie strengths. The experimental results show that the methods are effective.

CHAPTER 5

EVENT DETECTION

5.1. Introduction

Another piece of important information that can be extracted from call records is the occurrence of events. This could be used for detecting network attacks. In order to identify events in the call records I first used wavelet de-noising method to process the data and then applied the modified method in [Cangussu and Baron 2006] for detecting change points based on number of calls and call durations. These steps were described next.

Social network structures and relationships dynamically changes over time. Change point and event detection methods can be used to discover human relationship and behavior changes based on human communication pattern changes.

Change point detection is performed on a series of time ordered data in order to detect whether any change has occurred. It determines the number of changes and estimates the time of each change. Change point detection problems have many applications, including industrial quality control, reliability, fault detection, clinical trials, finance, environment, climate, signal detection, surveillance and security systems. Analyzing pattern changes of human behavior is an area of increasing interest in a number of different applications. The automatic detection of change points by studying patterns of human behavior is one of them and has recently attracted attention. One of the important applications of change point detection is in the area of homeland security. For example, terrorists or robbers usually conduct some attack in groups. The leader may communicate with his followers by wireless

phones for planning, coordinating, and executing the attacks. During this period of time their calling patterns are different from their usual ones. There are more calls or longer talk time than those of usual cases among their members, especially the leader. Also they meet at some particular place, especially the attack target place for planning and attacking. Combined with other evidences change point detection of calling patterns can be very useful to prevent security threats.

A lot of work on change point detection has been done previously. Baron (2000) proposed efficient on-line and off-line nonparametric algorithms for detecting the change-point based on histogram density estimators. Baron and Granott (2003) developed schemes for detecting early change points and frequent change points. They possessed a number of desired properties including distribution-consistency which implies convergence of small-sample change-point estimators. The above methods were applied to the temperatures, climate and software engineering quality control data (Cangussu and Baron 2006). Raftery and Akman (1986) developed a Bayesian approach for detecting a single change point at an unknown time of a Poisson process. Yang and Kuo (2001) proposed a Bayesian binary segmentation procedure for detecting multiple change points for a Poisson process. Ritov et al. (2002) applied the Bayesian change point method to neural data under the assumption of inhomogeneous Poisson process. Carlin et al. (1992) proposed the changing linear regression model and obtained the desired posterior densities by iterative Monte Carlo method. Lund and Reeves (2002) proposed the revision of two-phase linear regression model and applied it to climate data. Hawkins (1977) developed the procedure for detecting change points of a series of varying normal means under the assumption of the known variance. Worsley (1979) extended the method in (Hawkins 1977) for a

series of varying normal mean with unknown variance. These procedures were performed based on the likelihood ratio test. Chen and Gupta (1997) used a binary procedure combined with Swarz information criterion for testing and locating variance change points in a series of independent normal random variables under the assumption of the known and common mean. Johnson et al. (2003) applied reversible jump Markov chain Monte Carlo simulation to estimate variance change points of activation patterns from electromyographic data with assumption that data had a zero-mean. The variance is modeled by a step function. Chu and Zhao (2004) applied Bayesian approach to detect change points in the time series of annual tropical cyclone counts under assumption of a Poisson process with gamma distribution. Fearnhead (2006) performed direct simulation from the posterior distribution of multiple change point models with the unknown number of change points based on the recursions. The class of models assumes that the parameters associated with segments of data between successive change points are independent of each other. Yamanishi and Takeuchi (2006) proposed a scheme for detecting outliers and change-points for non-stationary time series. The main feature of this scheme is that the outlier is first detected by the model learned in the first stage which repeats the learning process twice and change point is detected by the learned model in the second one. They applied the scheme to AR models and SDAR algorithm as learning modules which are variants of maximum-likelihood method for online discounting learning of that model, which is adaptive to non-stationary time series. Erdman and Emerson (2008) applied the Bayesian method for the analysis of change points for the segmentation of micro-array data with implementation of the Bayesian change point method in linear time.

I used change point to refer to a large scale activity that changes relative to normal patterns of behavior. To understand such data, we often care about both the patterns of the typical behavior, and the detection and extraction of information from the deviations from this behavior.

I combined the wavelet denoising and sequential detection methods to detect change points based on call detail records. Call detail records did not document the conversation content. I can only use the information such as the time of the initiation of a call, number of call within a period of time, call duration, incoming call, outgoing call, and location.

5.2 Wavelet Overview

The transform of a signal is just another form of representing the signal. It does not change the information content presented in the signal. Fourier transform based signal analysis is used for frequency domain analysis. However, Fourier transform cannot provide any information of the signal changes with respect to time. Fourier transform assumes the signal is stationary, but most of signals in real world are always non-stationary. To overcome these shortcomings, the Short Time Fourier Transform (STFT) (Gabor 1946) was used to represent the signal in both time and frequency domain through a time window function which is nonzero for only a short period of time, which can be also used to analyze non-stationary signals. The window length determines a constant time and frequency resolution. Thus, a shorter time window is used in order to capture the transient behavior of a signal. STFT gives a constant resolution at all frequencies. A wide window gives better frequency resolution but poor time resolution. A narrower window gives good time resolution but poor frequency resolution. To overcome the deficiency of the STFT, the wavelet

transform was developed to provide a time-frequency representation of the signal. The wavelet transform uses multi-resolution technique by which different frequencies are analyzed with different resolutions.

Wavelet is associated with building a model for a signal, system or process with a set of special signals. Wavelet is localized waves which are oscillating function of time or space and are periodic. It has its energy concentrated in time or space and is suited to analysis of transient signals. While Fourier Transform and STFT use waves to analyze signals, the Wavelet Transform uses wavelets of finite energy.

The wavelet transform provides a time-frequency representation of the signal. The wavelet transform uses multi-resolution technique by which different frequencies are analyzed with different resolutions. In wavelet analysis the signal to be analyzed is multiplied with a wavelet function and the transform is computed for each segment generated. The wavelet transform at high frequencies gives good time resolution and poor frequency resolution, while at low frequencies, the wavelet transform gives good frequency resolution and poor time resolution. Sets of wavelets are employed to approximate a signal or process and each element in the wavelet set is constructed from the same function, the original wavelet, called the mother wavelet.

A generalized wavelet in normalized form is defined by

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}}\psi(\frac{t-b}{a}) \qquad (5.1)$$

where $a, b \in R, a > 0$, a is scale parameter, b is translation parameter and satisfies:

$$\int_{-\infty}^{\infty} \psi(t)dt = 0 \text{ and } \int_{-\infty}^{\infty} |\psi(t)|^2 \, dt < \infty$$

The wavelet transform calculates wavelet coefficient. The Continuous Wavelet Transform (CWT) is given by

$$W_{a,b} = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} x(t)\psi(\frac{t-b}{a})dt \qquad (5.2)$$

where $W_{a,b}$ is wavelet coefficients, *x(t)* is the signal to be transformed, $\psi(t)$ is the mother wavelet or basis function.

The scale parameter, *a*, indicates the level of analysis defined as 1/frequency and corresponds to frequency information. Scaling either dilates or compresses a signal. Large scales (low frequencies) dilate the signal and provide detailed information hidden in the signal, while small scales (high frequencies) compress the signal and provide global information about the signal. The multiplication of $\frac{1}{\sqrt{a}}$ is for energy normalization purposes so that the transformed signal will have the same energy at every scale. The translation parameter *b* relates to the location of the wavelet function as it is shifted through the signal. Thus, it corresponds to the time information in the wavelet transform. The wavelet transform decomposes the signal into different scales with different levels of resolution by dilating the mother wavelet. Furthermore, a mother wavelet has to have a zero net area, which suggests that the transformation kernel of the wavelet transform is a compact support function.

The wavelet coefficients are measures of the goodness of fit between the signal and the wavelet. Large coefficients indicate a good fit.

The inverse transform of CWT used to compute original data is given as:

$$x(t) = \sum_a \sum_b W_{a,b}\psi_{a,b}(t) + \sum_a c_{a,b}\phi_{a,b}(t) \qquad (5.3)$$

where $\phi_{a,b}(t)$ denotes the scaling function, $c_{a,b}$ denotes scaling coefficients which are defined by

$$c_{a,b} = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} x(t)\phi(\frac{t-b}{a})dt \qquad (5.4)$$

One drawback of the CWT is that the representation of the signal is often redundant, since a and b are continuous over the real number.

In Discrete Wavelet Transform (DWT) the scale factors between levels are usually chosen to be powers of 2. Widely used a and b parameters are set to $a = 2^j$ and $b = 2^j k$, $j, k \in Z$. For DWT the mother wavelet is defined as:

$$\psi_{j,k}(t) = 2^{-j/2}\psi(2^{-j}t - k) \qquad (5.5)$$

The DWT is given by

$$W_{j.k} = 2^{-j/2} \int_{-\infty}^{\infty} x(t)\psi(2^{-j}t - k)dt \qquad (5.6)$$

where $W_{j,k}$ is wavelet coefficients, x(t) is the signal to be transformed, ψ(t) is the mother wavelet or basis function.

A scaling function $\phi_{j,k}(t)$ (father wavelet) is introduced such that for each fixed j, the family

$$\phi_{j,k}(t) = 2^{-j/2}\phi(2^{-j}t - k) \qquad (5.7)$$

where $j, k \in Z$ and

$$\int \phi(t)dt = 1$$

The inverse transform of DWT used to compute original data is given as:

$$x(t) = \sum_{j=1}^{\infty} \sum_{k \in Z} W_{j,k}\psi_{j,k}(t) + \sum_{j=1}^{\infty} c_{j,k}\phi_{j,k}(t) \qquad (5.8)$$

where $\phi_{j,k}(t)$ denotes the scaling function, $c_{j,k}$ denotes scaling coefficients which are defined by

$$c_{j,k} = 2^{-j/2} \int_{-\infty}^{\infty} x(t) \phi_{j,k}(2^{-j} t - k) dt \qquad (5.9)$$

Given the discrete time signals *x[n] n ϵ Z*, where scale, translation and time are all discrete, by the discrete wavelet transform on L levels labeled from *j=1, ..., L*, the decomposition of *x[n]* given by

$$x[n] = \sum_{j=1}^{L} \sum_{k \in Z} W_{j,k} \psi_j [2^{-j} n - k] + \sum_{j=1}^{L} \sum_{k \in Z} c_{j,k} \phi [2^{-j} n - k] \qquad (5.10)$$

where the $\psi_j [2^{-j} n - k]$ is the discrete wavelet and $\phi_j [2^{-j} n - k]$ are the scaling sequences. The DWT computes the wavelet coefficients $W_{j,k}$ for j=1, …, L, and scaling coefficients $c_{j,k}$ by

$$W_{j,k} = 2^{-j} \sum_n x[n] \psi_j [2^{-j} n - k] \qquad (5.11)$$

and

$$c_{j,k} = 2^{-j} \sum_n x[n] \phi_j [2^{-j} n - k] \qquad (5.12)$$

The inverse wavelet transform (*IWT*) reconstructs the signal from its coefficients.

The practical usefulness of DWT comes from its Multi-Resolution Analysis (MRA) ability (Jawerth and Sweldens 1994), and efficient reconstruction filterbank structures. The fundamental concept involved in MRA is to find the average features and the details of the signal via scalar products with scaling signals and wavelets. The *MRA* decomposes a signal into scales with different time and frequency resolutions. *MRA* is designed to give good time resolution and poor frequency resolution at high frequencies and good frequency resolution and poor time resolution at low frequencies.

In DWT a time-scale representation of the signal is obtained using filtering techniques. Filter is one of the most widely used signal processing functions. The signal to be analyzed is passed through filters with different cutoff frequencies at different scales. Wavelets can be realized by iteration of filters with rescaling. The resolution of the signal, which is a measure of the amount of detail information in the signal, is determined by the filtering operations, and the scale is determined by upsampling and downsampling operations.

The DWT is computed by successive low-pass and high-pass filtering of the discrete time-domain signal. Assuming that the sequence x_n denotes the signal, where n is an integer, the *h* denotes high-pass filter and the *g* denotes low-pass filter. At each level, the high-pass filter associated with wavelet function produces detailed coefficient, d_n, while the low-pass filter associated with scaling function produces approximated coefficient, a_n.

By MRA principal in (Jawerth and Sweldens 1994), the scaling function satisfies the following 2-scale (dilation or refinement) equation

$$\phi(t) = \sum_{n=-\infty}^{\infty} g[n]\phi[2t-n] \tag{13}$$

where $n \epsilon Z$ and it satisfies the condition $\sum_n g[n] = 1$

The wavelet function satisfies equation

$$\psi(t) = \sum_{n=-\infty}^{\infty} h[n]\phi[2t-n] \tag{14}$$

where $n \epsilon Z$ with the conditions $\sum_n h[n] = 0$ and $h[n] = (-1^n)g(-n+1)$

These equations can be implemented as a tree-structured filter-bank shown in figure 5.1. At each decomposition level, the half band filters produce signals spanning

over only half the frequency band. This doubles the frequency resolution as the uncertainty in frequency is reduced by half. The process continues until the desired level is reached. The maximum number of levels depends on the length of the signal. The *DWT* of the original signal is then obtained by concatenating all the coefficients, starting from the last level of decomposition.

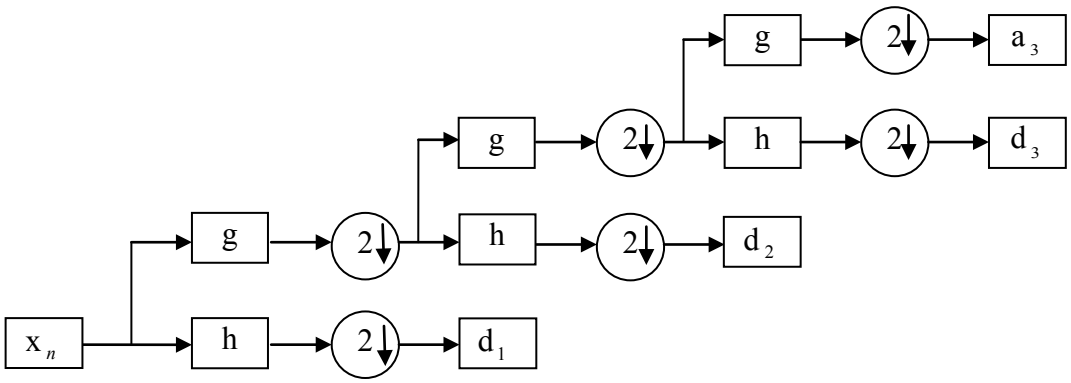


Fig. 5.1 Three-level wavelet decomposition.

The reconstruction of filter-bank structure is shown in figure 5.2. The reconstruction is the reverse process of decomposition. The wavelet coefficients and scaling coefficient at every level are upsampled by two, passed through the high pass *h'* and low pass *g'* synthesis filters and then added. The decomposition and synthesis filters are identical to each other, except for a time reversal. This process continues through the same number of levels as in the decomposition process to obtain the original signal.

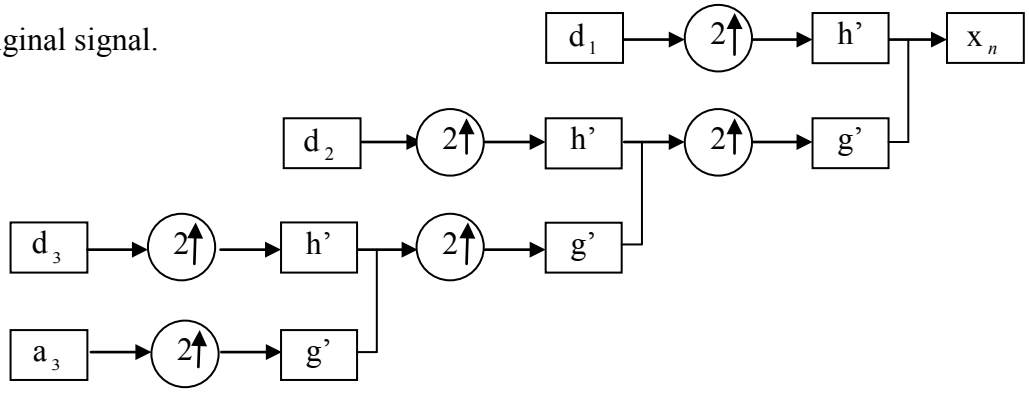


Fig. 5.2 Three-level wavelet reconstruction.

There are a number of basic functions that can be used as the mother wavelet for wavelet transformation. Since the mother wavelet produces all wavelet functions used in the transformation through translation and scaling, it determines the characteristics of the resulting wavelet transform. Therefore, the details of the particular application should be taken into account and the appropriate mother wavelet should be chosen in order to use the wavelet transform effectively.

Haar wavelet is one of the simplest wavelet. The Haar scaling function is defined as

$$\phi(t) = \begin{cases} 1 & if\ 0 \le t < 1 \\ 0 & otherwise \end{cases} \tag{5.15}$$

where $\phi(t)$, the Haar scaling function, is a step function. The scaling function can be described with a dilation equation of the type

$$\phi(t) = \sum_k c_k \phi[2t - k] \tag{5.16}$$

where the constants c_k are the refinement coefficients. For the Haar scaling function,

Equation 5.16 can be written as

$$\phi(t) = \phi(2t) + \phi(2t - 1) \tag{5.17}$$

where the refinement coefficients $c_0 = c_1 = 1$ and all others are zeros.

The Haar mother wavelet is defined as

$$\psi(t) = \begin{cases} 1 & if\ 0 \le t < \frac{1}{2} \\ -1 & if\ \frac{1}{2} \le t < 1 \\ 0 & otherwise \end{cases} \tag{5.18}$$

The Haar mother wavelet is constant on intervals of one-half and is related to the scaling function by

$$\psi(t) = \phi(2t) - \phi(2t-1) \qquad (5.19)$$

For the Haar wavelet transformation, consider a signal x_n of 2^n sample values $x_{n,l}$:

$$x_n = \{x_{n,l} \mid 0 \le l < 2^n\}$$

Apply the average and difference transform onto each pair x_{2l} and x_{2l+1}. There are 2^{n-1} such pares (l=0, 1, …, 2^{n-1}). The average and difference are denoted by $x_{n-1,l}$ and $d_{n-1,l}$:

$$x_{n-1,l} = \frac{x_{n,2l} + x_{n,2l+1}}{2}$$

$$d_{n-1,l} = \frac{x_{n,2l} - x_{n,2l+1}}{2}$$

The input signal x_n which has 2^n sample is split into two signals x_{n-1} with 2^{n-1} averages $x_{n-1,l}$ and d_{n-1} with 2^{n-1} differences $d_{n-1,l}$. We have not lost any information because given the averages $x_{n-1,l}$ and differences $d_{n-1,l}$ we can always recover $x_{n,2l}$ and $x_{n,2l+1}$ as

$$x_{n,2l} = x_{n-1,l} + d_{n-1,l}$$

$$x_{n,2l+1} = x_{n-1,l} - d_{n-1,l}$$

Therefore, given averages x_{n-1} and differences d_{n-1}, one can reconstruct the original signal x_n.

As an example to show the Haar wavelet transformation, consider a signal that has been sampled on intervals of length 1/4 over the range [0, 1] so as to produce the sequence $\{8,4,1,3\}$. The low-pass and the high-pass of the Haar wavelet decomposition are the average and difference transform respectively. To find the wavelet coefficients, the decomposition is performed on a series of averages and differences of sequential pairs of elements in the signal, as indicated in Figure 5.3. The wavelet decomposition result is $\{4,2,2,-1\}$. The Harr wavelet reconstruction process of the sequence $\{8,4,1,3\}$ is shown in Figure 5.4.

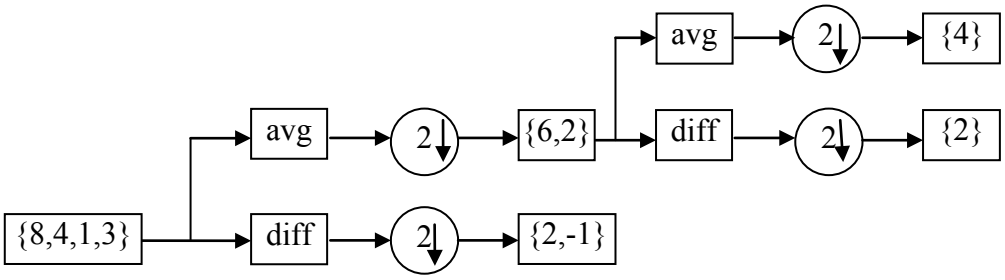


Fig. 5.3 Harr wavelet decomposition by averaging and differencing.

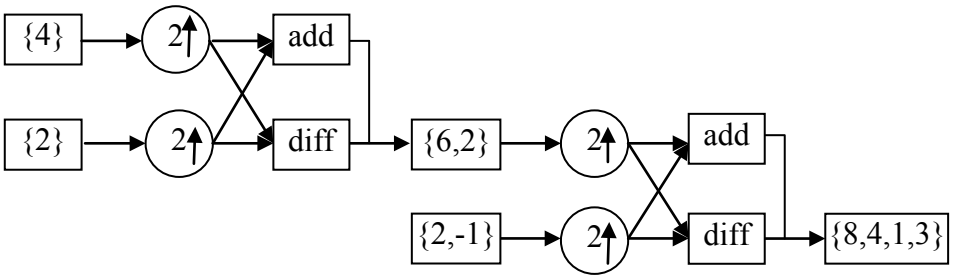


Fig. 5.4 Harr wavelet reconstruction.

5.3 Wavelet Denoising

Generally, for the denoising, the wavelet scaling function should have properties similar to the original signal. The general wavelet de-noising procedure follows 3 steps:

1. Wavelet Selection

The differences among different mother wavelet functions (e.g. Haar, Daubechies, Coiflets, Symlet, Biorthogonal) exist in how these scaling signals and the wavelets are defined. The choice of wavelet determines the final waveform shape. To best characterize the change points in a noisy signal, one should select the wavelet to better approximate and capture the transient change points of the original signal. The choice of mother wavelet can be based on correlation between the signal of interest and the wavelet denoised signal given by

$$\gamma = \frac{\sum (X - \overline{X})(Y - \overline{Y})}{\sqrt{\sum (X - \overline{X})^2 (Y - \overline{Y})^2}}$$

where $\overline{X}$ and $\overline{Y}$ are the mean value of set of the signal of interest X and the wavelet denoised signal set Y, respectively.

2. Threshold Selection

After a suitable wavelet basis function is chosen, the *DWT* decomposition of a signal will compress the energy of the signal into a small number of large magnitude wavelet coefficients. The *DWT* transforms the Gaussian white noise in any one orthogonal basis into wavelet coefficients of small magnitude. This property of the *DWT* allows the suppression of noise by applying a threshold which retains wavelet coefficients representing the signal and removes low magnitude coefficients which represent noise.

Assuming a finite observation signal of length n $x = (x_1, x_2, ..., x_n)$ is a noisy data of the signal $s = (s_1, s_2, ..., s_n)$, then

$$x_i = s_i + \varepsilon_i$$

for i=1, …, n, where ε_i is noise.

The waveShrink method (Donoho and Johnstone 1994) is widely used to estimate signal *x*. The two commonly used shrinkage functions are hard and soft thresholding functions defined as

$$S_\delta^{hard}(x) = \begin{cases} x & |x| > \delta \\ 0 & |x| \le \delta \end{cases}$$

$$S_\delta^{soft}(x) = \begin{cases} x-\delta & x > \delta \\ \delta - x & x < -\delta \\ 0 & |x| < \delta \end{cases}$$

where $\delta \ge 0$ is the threshold.

The threshold is usually determined in one of the following four ways.

Universal Threshold: The universal threshold is defined as (Donoho and Johnstone 1993):

$$\delta_U = \sigma\sqrt{2\log(n)}$$

where σ is standard deviation of the noise and n is the sample length. It uses a single threshold for all wavelet coefficients. One drawback is it does not work well for signals of short duration. Other methods may provide more accurate results.

SURE Threshold: The SURE threshold is based on Stern's Unbiased Risk Estimator (Stern 1981). This method minimizes the SURE function to determine an optimal threshold. The SURE threshold is defined as

$$\delta_{SURE} = \min_\delta SURE(\delta, \frac{x}{\sigma})$$

where SURE() is defined as

$$SURE(\delta, X) = n - 2m_{\{i:|X_i| \le \delta\}} + \sum_{i=1}^{n} [\min(|X_i|, \delta)]^2$$

where δ is the candidate threshold, x_i is the wavelet coefficient , n is the data size, and m is the number of the data point less than δ. This method has the sparse wavelet coefficient problem and it is often combined with universal threshold in a hybrid method.

Hybrid Threshold*:* The hybrid threshold method is a combination of the universal and SURE threshold methods (Donoho and Johnstone 1993). This method uses the universal threshold if the signal to noise ratio (SNR) is low with sparse wavelet coefficient, otherwise uses SURE threshold method.

Minmax Threshold*:* The Minmax threshold method uses a fixed threshold selected to produce minmax performance for the mean square error (MSE). The Minmax uses a single threshold for all wavelet coefficients. It is defined as (Donoho and Johnstone 1994)

$$\delta_{\min i\max} = \inf_{\delta} \sup_{\mu} \left\{ \frac{R_{\delta}(\mu)}{n^{-1} + \min(\mu^2, 1)} \right\}$$

where $R_{\delta}(\mu) = E(S_{\delta}(x) - \mu)^2$, $x \sim N(\mu, 1)$.

3. Inverse Wavelet Transform

The inverse wavelet transform (IWT) of DWT reconstructs the signal from its coefficients by formula 5.7.

5.4 Change Point Detection Procedures

Change point analysis method attempts to find a point along a distribution or trend of values where the characteristics of the values before and after the point are different.

Let $X = (X_1, X_2, ..., X_{\theta})$ be a process.

The multiple change points for the process are defined as (Cangussu and Baron 2006)

$$X_1 = (x_1,...,x_{\tau_1}) \sim f_1$$
$$X_2 = (x_{\tau_1+1},...,x_{\tau_2}) \sim f_2$$
$$.....$$
$$X_\theta = (x_{\tau_{\theta-1}+1},...,x_{\tau_\theta}) \sim f_\theta$$

where $f_1, f_2,..., f_\theta$ are either known or unknown probability density functions or trends; $\tau_1, \tau_2,..., \tau_{\theta-1}$ are change points.

I first used wavelet denoising method to pre-process the data, and then used sequential estimation scheme in [Cangussu and Baron 2006] for detecting multiple change points which chooses increasing subsamples and finds one change point at a time until all change points are found.

The general procedure for change point detection is as follows.

1. Sequential Detection

A widely used change point detection method based on Page's cumulative sum (cusum) rule is defined by

$$T(a) = \inf\{n : S_n \geq a\}$$

where

$$S_n = \max_{1 \leq k \leq n} \sum_{i=k+1}^{n} \log \frac{g(x_i)}{f(x_i)}$$

is maximum likelihood ratio based cumulative sum, *a* is a threshold, f and g are pre- and post-change density functions. When the density functions are unknown, the best estimates for each value k of a change point are used, the cusum is computed by

$$\hat{S}_n = \max_{1 \leq k \leq n} \sum_{i=k+1}^{n} \log \frac{\hat{g}(x_i)}{\hat{f}(x_i)}$$

and the stopping rule is defined by (Cangussu and Baron 2006)

$$\hat{T}(a) = \inf\{n : \hat{S}_n \geq a\}$$

The call log data is not a classical change point model type. I used the method by Cangussu and Baron (2006) to detect trend changes, either linear or exponential, in each segment between change points. The computation of the above cusum change point detection method will be more complicated if there are unknown parameters in segments (Cangussu and Baron 2006).

I used least-squares method for the nonlinear trends. The trend which is linear or exponential is decided by the lower sum of squares for each segment. In this way the maximum likelihood ratio is replaced by the minimum weighted sum of squared residuals

$$\hat{S}_n = \min_k\{\sum_{i=1}^{k}\sqrt{\hat{f}_i}(x_i - \hat{f}_i) + \sum_{i=k+1}^{n}\sqrt{\hat{g}_i}(x_i - \hat{g}_i)\}$$

and the stopping rule is defined as

$$\hat{T}(\alpha) = \inf\{n : p_n \leq \alpha\}$$

where p_n is a p-value testing significant based on $\hat{S}_n$ of a change point at k.

2. Post-estimation

The detected change point must be estimated after it is detected by a stopping rule. The change point estimator I used is based on the cusum stopping rule $\hat{T}$ and the minimum p-value

$$\hat{\tau} = \arg\min_{1\leq k<\hat{T}} p(k, \hat{T}, X)$$

where $p(k, \hat{T}, X)$ is the p-value of the likelihood ratio test comparing $X_1 = (x_{1,\ldots,} x_k)$ and $X_2 = (x_{k+1,\ldots,} x_{\hat{T}})$.

3. Significance Tests

To eliminate false change point, I used ANOVA F-type tests. If the test is significant, I repeated step 1-3 to search for next change point, else, it is a false change point and the search continues based on initial sequence after the last significant change point.

For fitting linear trend $E(x_i) = a + bt_i$, I used the standard least square estimates

$$\hat{b} = \frac{\sum_i (x_i - \bar{x})(t_i - \bar{t})}{\sum_i (t_i - \bar{t})^2} \quad \text{and} \quad \hat{a} = \bar{x} - \hat{b}\bar{t}$$

and for fitting exponential trend $E(x_i) = \exp(a + bt_i) - 1$, I used

$$\hat{b} = \frac{\sum_i \log(1 + x_i)(t_i - \bar{t})}{\sum_i (t_i - \bar{t})} \quad \text{and} \quad \hat{a} = \overline{\log(1+x)} - \hat{b}\bar{t}$$

for initial approximation.

When the preliminary estimator of a change point was obtained, a refinement of this estimator was performed by least square fitting from the segment in the neighborhood of the preliminary estimator. If the change points were not significant for the chosen level α, they were removed and the corresponding segments were merged.

After the iterations end, all the change points are significant at the chosen level α.

I selected Coiflets5 wavelet and Minimax threshold method to denoise the data by the principles described above, and then applied the sequential change point detection method.

I used both simulation data and the real data from the data sets.

The simulation data sets are randomly generated based on

$X = (X_1, X_2, ..., X_\theta)$, $X_i \sim N(\mu_i, \sigma_i^2)$ for i=1, 2, …, θ.

5.5 Unusual Consumption Event Detection

Almost all previous work on event detection is based on text, data stream and video. Brants and Chen (2003) proposed a method based on an incremental term frequency-inverse document frequency (TF-IDF) model and the extensions include generation of source-specific models, similarity score normalization based on document-specific averages, source-pair specific averages, term re-weighting based on inverse event frequencies, and segmentation of the documents. Chen et al. (2003) examined the effect of a number of techniques, such as part of speech tagging, similarity measures, an expanded stop list on the performance. Kumaran and Allan (2004) used text classification techniques and named entities to improve the performance. Li and Croft (2005) proposed a novelty detection approach based on the identification of sentence level patterns. Li et al. (2005) proposed a probabilistic model to incorporate both content and time information in a unified framework which gave new representations of both news articles and news events. They explored two directions because the news articles are always aroused by events, and similar articles reporting the same event often redundantly appear on many news sources. Zhao and Mitra (2007) proposed a method to detect events by combining text-based clustering, temporal segmentation, and graph cuts of social networks in which each node represents a social actor and each edge represents a piece of text communication that connects two actors. He et al. (2006) proposed the conceptual model-based approach by the use of domain knowledge and named entity type assignments and showed that classical cosine similarity method fails in the anticipatory event detection task. Luo et al. (2007) proposed the online new event detection (ONED) framework. They

combined the indexing and compression methods to improve the document processing rate. They applied a resource-adaptive computation method to maximize the benefit that can be gained from limited resources. The new events are further filtered and prioritized before they are presented to the consumer when the new event arrival rate is beyond the processing capability of the consumer. The implicit citation relationships are created among all the documents and used to compute the importance of document sources.

Guralnik and Srivastava (1999) proposed an iterative algorithm and used likelihood criterion to segment a time-series into piecewise homogeneous regions to detect the change points which are equivalent to events defined by the authors and evaluate them with the highway traffic data. Kleinberg (2002) used an infinite automaton in which bursts are state transitions for detecting burst events in text streams and conducted the experiments with emails and research papers. Keogh et al. (2002) used suffix tree to encode the frequency of all observed patterns and applied a Markov model to detect patterns in the symbol sequence. Salmenkivi and Mannila (2005) found piecewise constant intensity functions which can be used to represent continuous intensity functions using a combination of Poisson models and Bayesian estimation methods. They used dynamic programming methods to find bursts. Ihler et al. (2006) used a time-varying Poisson process model and statistical estimation techniques for unsupervised learning in the context. They applied this model to freeway traffic data and building access data.

I proposed and investigated the inhomogeneous Poisson and inhomogeneous exponential distribution model to detect events, and I illustrated how to learn such a model from data to both characterize normal behavior and detect anomalous events

based on call detail records. There is no conversation content in call detail records that is the main difference between my data and the text and website data. Therefore it is more difficult to detect events hidden in them. I can only used information such as the time of the initiation of calls, number of calls within a period of time, call duration, incoming calls, outgoing calls and location. The maximum likelihood estimation (Harris and Stocker 1998) was used to estimate the rates and the thresholds of the number of calls and call duration.

Change point detection methods do not deal with bursts of short width in time series. The bursts in time series are related to some events such as attacks in networks. I proposed the inhomogeneous Poisson model for detecting the bursts, which I labeled them unusual consumption events.

Assuming that number of calls follows inhomogeneous Poisson process and call duration follows inhomogeneous exponential distribution.

Let $N_i = \{n_{i1}, \ldots n_{ik}\}$ be random variable for number of calls of a given day i, $D_i = \{d_{i1}, \ldots, d_{ik}\}$ be random variable for call duration for day i, $i=1, 2, \ldots, 7$ be a day of week, 1 for Sunday, …, 7 for Saturday. Then

$$N = \begin{bmatrix} n_{11} & n_{12} & \ldots & n_{1k} \\ n_{21} & n_{22} & \ldots & n_{2k} \\ \ldots & \ldots & \ldots & \ldots \\ n_{71} & n_{72} & \ldots & n_{7k} \end{bmatrix}$$

is the matrix of number of calls on 7 days of week and

$$D = \begin{bmatrix} d_{11} & d_{12} & \ldots & d_{1k} \\ d_{21} & d_{22} & \ldots & d_{2k} \\ \ldots & \ldots & \ldots & \ldots \\ d_{71} & d_{72} & \ldots & d_{7k} \end{bmatrix}$$

is the matrix of call duration on 7 days of week.

Then Poisson density function for day *i* is given by

$$P_{N_i}(N_i = n_{ij}) = \frac{e^{-\lambda_i}\lambda_i^{n_{ij}}}{n_{ij}!}$$

where λ_i is the rate (average) of number of calls for day *i*. By the properties of Poisson distribution, the mean= λ_i , the variance var= λ_i and the standard error $\sigma = \pm\sqrt{\lambda_i}$.

The exponential distribution density function of call duration for day *i* is given by

$$P_{D_i}(D_i = d_{ij}) = \frac{1}{\mu_i} e^{-\frac{d_{ij}}{\mu_i}}$$

where μ_i is the mean of call duration for day *i*.

By the properties of exponential distribution, the variance var= μ_i^2 and the standard error $\delta = \pm\sqrt{\mu_i^2} = \pm\mu_i$

Now maximum likelihood [Harris and Stocker 1998] is used to estimate λ_i for day *i*. The cumulated probability distribution function is

$$P_{N_i}(N_i = n_{i1}, n_{i2}, ..., n_{ik} \mid \lambda_i) = \prod_{j=1}^{k} \frac{e^{-\lambda_i}\lambda_i^{n_{ij}}}{n_{ij}!}$$

$$= \frac{e^{-k\lambda_i}\lambda_i^{\sum_{j=1}^{k} n_{ij}}}{\prod_{j=1}^{k} n_{ij}!}$$

$$\ln P_{N_i} = -k\lambda_i + (\ln \lambda_i)\sum_{j=1}^{k} n_{ij} - \ln(\prod_{j=1}^{k} n_{ij})$$

$$\frac{d(\ln P_{N_i})}{d\lambda_i} = -k + \frac{\sum_{j=1}^{k} n_{ij}}{\lambda_i} = 0$$

$$\hat{\lambda}_i = \frac{\sum_{j=1}^{k} n_{ij}}{k} \tag{5.12}$$

For μ_i the cumulated probability distribution function of call duration is

$$P_{D_i}(D_i = d_{i1}, d_{i2}, ..., d_{ik} \mid \mu_i) = \prod_{j=1}^{k} \frac{1}{\mu_i} e^{-\frac{d_{ij}}{\mu_i}}$$

$$= \frac{1}{\mu_i^k} e^{-\frac{1}{\mu_i}\sum_{j=1}^{k} d_{ij}}$$

$$\ln P_{D_i} = -k \ln \mu_i - \frac{1}{\mu_i} \sum_{j=1}^{k} d_{ij}$$

$$\frac{d(\ln P_{D_i})}{d\mu_i} = -\frac{k}{\mu_i} + \frac{1}{\mu_i^2} \sum_{j=1}^{k} d_{ij} = 0$$

$$\hat{\mu}_i = \frac{\sum_{j=1}^{k} d_{ij}}{k} \tag{5.13}$$

The maximum likelihood estimates are used to estimate average number of calls and call duration. Next I considered the maximum average number of calls and call duration obtained for all weekday/weekend and week by week. Suppose that the m week data is used to compute the rates of number of calls and call duration for user p. Let $\hat{\lambda}_{d1}^{p}, \hat{\lambda}_{d2}^{p}, ..., \hat{\lambda}_{d7}^{p}$ be the rate of number of calls obtained for all weekday/weekend and $\hat{\lambda}_{w1}^{p}, \hat{\lambda}_{w2}^{p}, ..., \hat{\lambda}_{wm}^{p}$ be the rate of call duration obtained week by week for m weeks of user p respectively. Let $\hat{\mu}_{d1}^{p}, \hat{\mu}_{d2}^{p}, ..., \hat{\mu}_{d7}^{p}$ be the mean of call duration obtained for all weekday/weekend and $\hat{\mu}_{w1}^{p}, \hat{\mu}_{w2}^{p}, ..., \hat{\mu}_{wm}^{p}$ be the mean of call duration obtained week by week for m weeks of user p respectively.

Then the maximum means of number of calls and call duration are respectively computed by:

$$\hat{\lambda}_{\max}^{p} = \max(\hat{\lambda}_{d1}^{p}, \hat{\lambda}_{d2}^{p}, ..., \hat{\lambda}_{d7}^{p}, \hat{\lambda}_{w1}^{p}, \hat{\lambda}_{w2}^{p}, ..., \hat{\lambda}_{wm}^{p}) \tag{5.14}$$

$$\hat{\mu}^p_{\max} = \max(\hat{\mu}^p_{d1}, \hat{\mu}^p_{d2}, ..., \hat{\mu}^p_{d7}, \hat{\mu}^p_{w1}, \hat{\mu}^p_{w2}, ..., \hat{\mu}^p_{wm}) \quad (5.15)$$

where $\hat{\lambda}^p_{\max}$ and $\hat{\mu}^p_{\max}$ are the maximum likelihood estimates of number of calls and call duration for user p over the number of days specified respectively. The thresholds define the limits for all weekday/weekend and week by week. The assumption is that the calling pattern could be different. Each person has his/her own thresholds, and if the number of calls or call duration is greater than the thresholds of their own for some day, I define that there is an event in that day.

To calculate the threshold of number of calls for user p, N^p_{thres}, I define

$$N^p_{thres} = \hat{\lambda}^p_{\max} + \hat{\sigma}^p_{\max} \quad (5.16)$$

where $\hat{\lambda}^p_{\max}$ and $\hat{\sigma}^p_{\max}$ are the maximum rate of number of calls and correspondent standard error with positive $\hat{\sigma}^p_{\max}$.

To calculate the threshold of call duration for user p, D^p_{thres}, I define

$$D^p_{thres} = \hat{\mu}^p_{\max} + \hat{\delta}^p_{\max} \quad (5.17)$$

where $\hat{\mu}^p_{\max}$ and $\hat{\delta}^p_{\max}$ are the maximum mean of call duration and correspondent standard error with positive $\hat{\delta}^p_{\max}$.

Definition of an unusual consumption event

A collection of call log data can be represented as $C = <(t_1, a_1, d_1, l_1), (t_2, a_2, d_2, l_2), ...(t_n, a_n, d_n, l_n)>$,

where t_i is a time point, d_i is a call duration, l_i is a location and a_i is a pair of actors, caller-callee $< s_i, r_i >$ where s_i is an actor who initiates a call at time t_i and r_i is an

actor who receive a call. An unusual consumption event is defined as a subset E $\subset$C of a tuple $E = \{(t_1, a_1, d_1, l_1), (t_2, a_2, d_2, l_2), ...(t_m, a_m, d_m, l_m)\}$ such that either $\sum_{i=1}^{m} d_i > D_{thres}$ or $count(d_i) > N_{thres}$ defined as the above in the time period $\Delta t = t_m - t_1$.

The formulas (5.16) and (5.17) to detect events are based on the day of week, call frequencies and call duration.

Day of week*:* Everyone has his/her own schedule for working, studying, entertainment, traveling and so on. The schedule is mainly based on the day of the week.

Call frequencies*:* The call frequency is the number of incoming or outgoing calls within a period of time. The greater the number of incoming or outgoing calls during a period of time, the more socially close the caller and callee relationship is.

Call duration*:* The call duration is how long both caller and callee want to talk with each other. The longer the call duration is during a period of time, the more socially close the caller and callee relationship is.

The thresholds of frequency and duration are computed by formulas (5.16) and (5.17) based on day of week (Sunday, Monday, … , Saturday) and week sequence (1st week, 2nd week, …). Then the thresholds of frequency and duration are chosen to compare with the frequency and duration for each day. If the frequency or duration of some day is greater than the thresholds of frequency or duration, I define that there is an event in that day.

5.6 Experimental results and Discussions

5.6.1 Change Point Detection

The evaluation of change point detection was conducted first using synthetic data generated by simulation and then using real data.

Many workers have weeks when they are very busy (I refer to them as “busy weeks”) and others where the work to be performed is less than the regular load (I refer to them as “easy weeks”). The expectation is that during the busy weeks the worker makes and receives lower number of calls and has shorter talk time than those of usual weeks. During the easy weeks there is a tendency that he/she will make and receive higher number of calls with longer talk time than those of usual weeks. I randomly generated multiple simulation data sets of number of calls and call duration for 120 days for simulation. Fig. 5.5 and 5.6 are examples of this behavior.

Fig. 5.5 and 5.6 show the change points for the number of calls and call duration of dataset1 and dataset2, where the x-axis indicates the days and y-axis indicates the number of calls and call duration (minutes) respectively. The blue color dotted curve and green color solid curve indicate original and denoised ones, respectively. The vertical lines indicate the change points. The three detected change points happened on the 56^{th}, 60^{th} and 84^{th} days which match the change points of the curves.

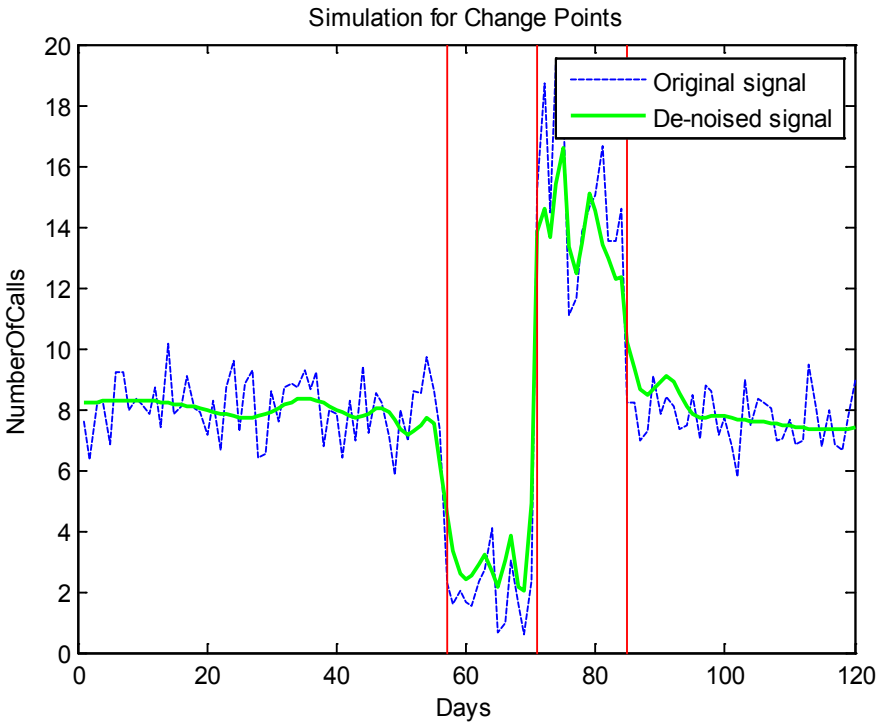


Fig. 5.5 The change points based on number of calls for simulation data.

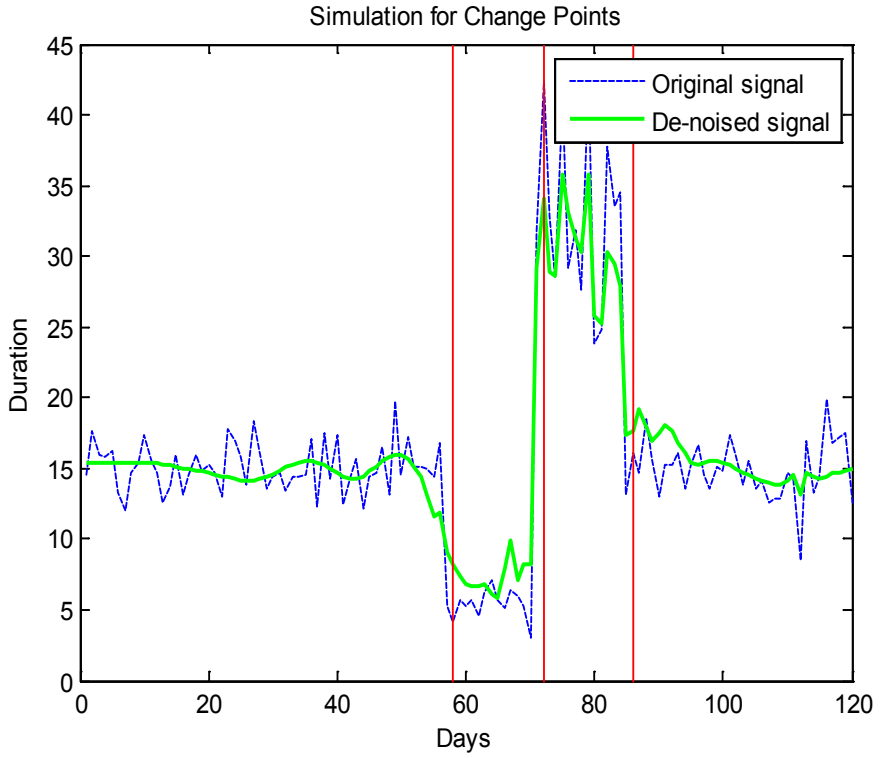


Fig. 5.6 The change points based on call duration for simulation data.

After multiple experiments in real data, I selected Coiflets5 wavelet and Minimax threshold method by the principles described above to denoise the data, and then applied the sequential change point detection method in (Cangussu and Baron 2006).

Fig. 5.7 and 5.8 show the change points for the number of calls and call duration of user 3, where the x-axis indicates the days and the y-axis indicates the number of calls and call duration respectively. The dotted curve and the solid curve indicate, respectively, original and de-noised data. The vertical lines indicate the change points. There are 3 change points identified at the 32^{nd}, 48^{th}, and 64^{th} day which correspond to Friday, Sunday, and Sunday respectively.

From the 1^{st} day to the 32^{nd} day, the user 3 visited New York City, world trade center and Harvard University. The average number of calls was 8 and the average duration was 7.5 minutes per day. Between the 32^{nd} and 48^{th} day, the user's activities were in local area. The average number of calls was 12 and the average duration was 10 minutes per day. From the 48^{th} day to the 64^{th} day, the user's activities were in local area. The average number of calls was 9 and the average duration was 6 minutes per day. From the 64^{th} day to the 108^{th} day, the user's activities were in local area. The average of number calls was 2 and the average duration was 2 minutes per day.

Results showed that most of change point days are associated with weekends when most people have some leisure time, which consequently changes their calling pattern. Although a sophisticated technique is not needed to find out change points over the weekends it should be noticed that my goal here is to show that my technique can identify change points based on call detail records. The actual change point can

represent other behaviors, for example, the individual under observation is a potential threat to public security.

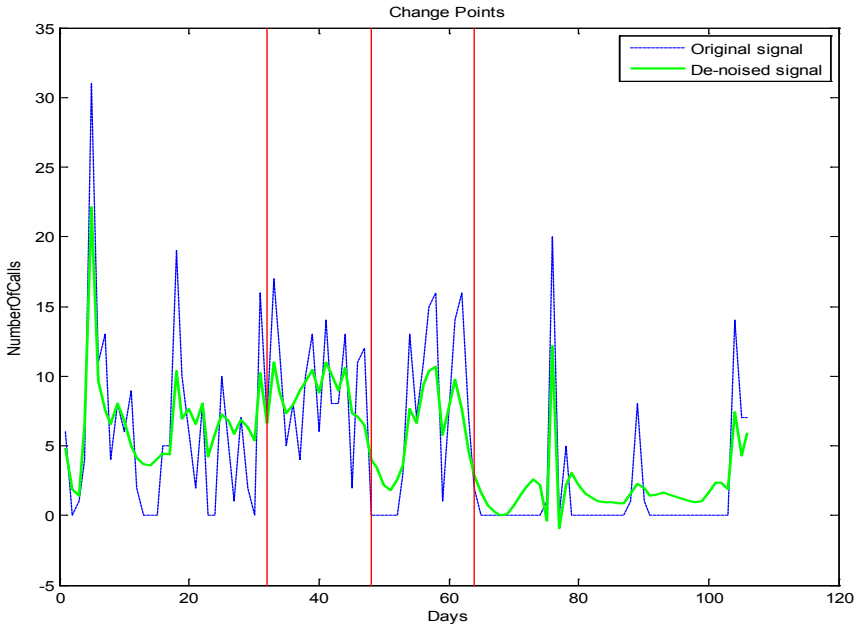


Fig. 5.7 The change points based on number of calls for user3.

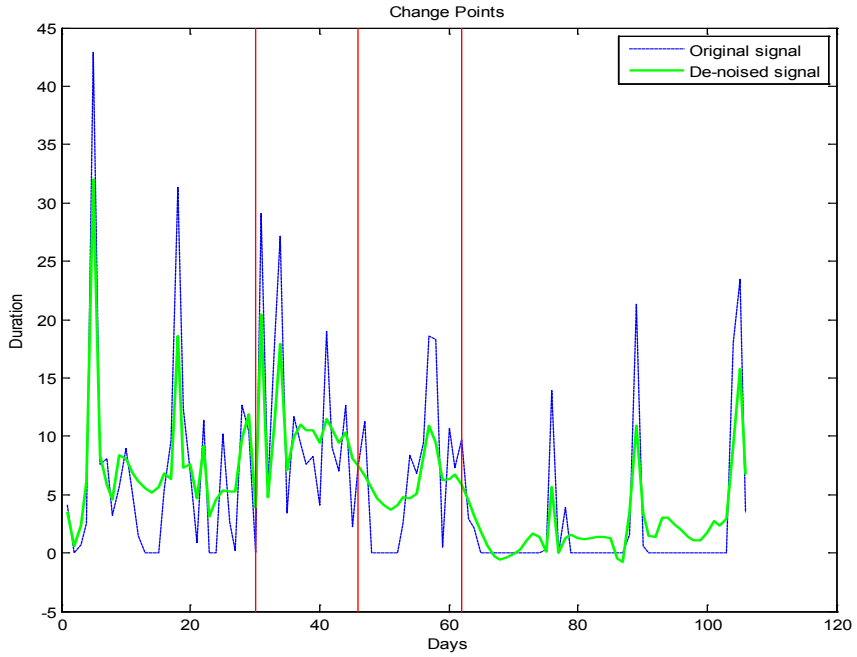


Fig. 5.8 The change points based on duration for user 3.

Fig. 5.9 and 5.10 show the change points for the number of calls and call duration of user 74, where the x-axis indicates the days and y-axis indicates the number of calls and call duration respectively. The blue color dotted curve and green color solid curve indicate original and de-noised ones, respectively. The vertical lines indicate the change points. In Fig. 5.9 there are 4 change points identified, the 21st, 41st, 61st and 91st day (last day of the fall semester) which are Tuesday, Saturday, Friday and Sunday respectively. In Fig. 5.10 there are 5 change points identified, the 21st, 41st, 64th, 86th and 107th day.

From the 1st day to the 21st day, user74 visited New York City at weekend. The average number of calls was 2 and the average duration was 1.5 minutes per day. Between the 21st and 41st day, he visited New York City at weekend. The average number of calls was 3.6 and the average duration was 11.8 minutes per day. From the 41st to 61st day, he visited New York City at Halloween week and weekend. The average number of calls was 2.5 and the average duration was 4.8 minutes per day. From the 61st day to the 91st day he visited Los Angelus, Providence, RI, and New York City at weekend. The average number of calls was 3 and the average duration was 8.5 minutes per day. From the 91st day to the 157st day, he visited New York City, Providence, RI, because of holidays. The average number of calls was 3.6 and the average duration was 6.7 minutes per day.

Results showed that most of change point days are weekends.

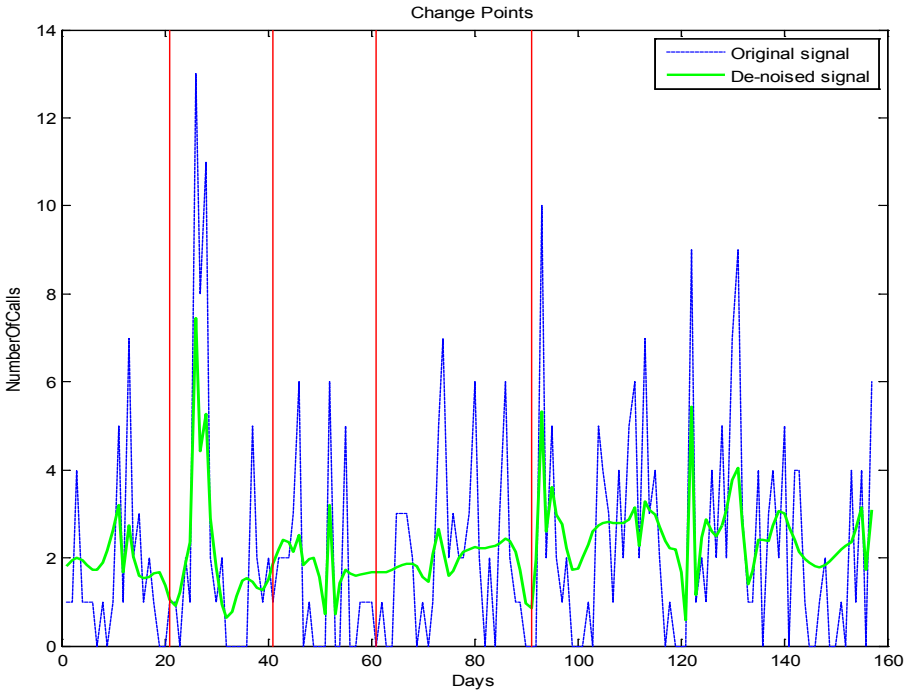


Fig. 5.9 The change points based on number of calls for user 74.

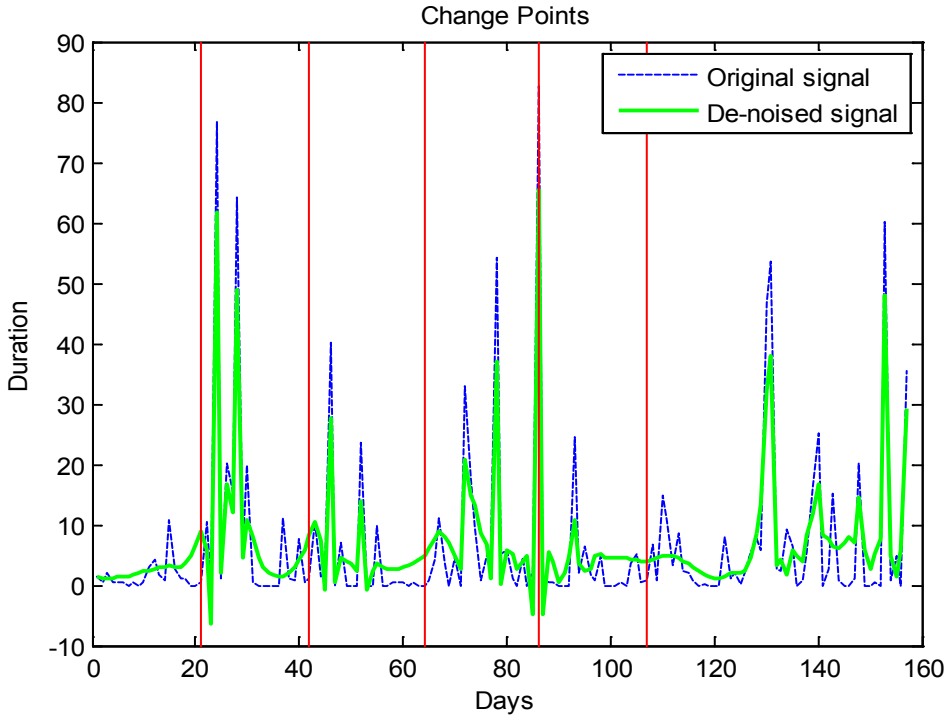


Fig. 5.10 The change points based on duration for user 74.

To evaluate the performance of the change point detection method, experiments were conducted on 20 phone users randomly chosen from actual call logs of 81 phone users. The significant level and the latency are two important measurements when considering the accuracy in the change point detection. The lower the significant level, the more sensitive the strategy is. This can avoid false positive. In all data sets I used a significant level of 0.01.

The shorter the latency, the better the indication of the accuracy of the strategy is. So a minimization of the latency is desired. In this study a minimum of 10 data points were chosen to start the detection of a new change point. A total of 81 change points were identified for 20 users. 78% of the change points were identified using about 12 data points and 10% using about 15 data points shown in Fig. 5.11. In only a very few cases, the change points were identified later than the 20^{th} points as shown in Fig. 5.11. This is a very good indication of the accuracy of the strategy.

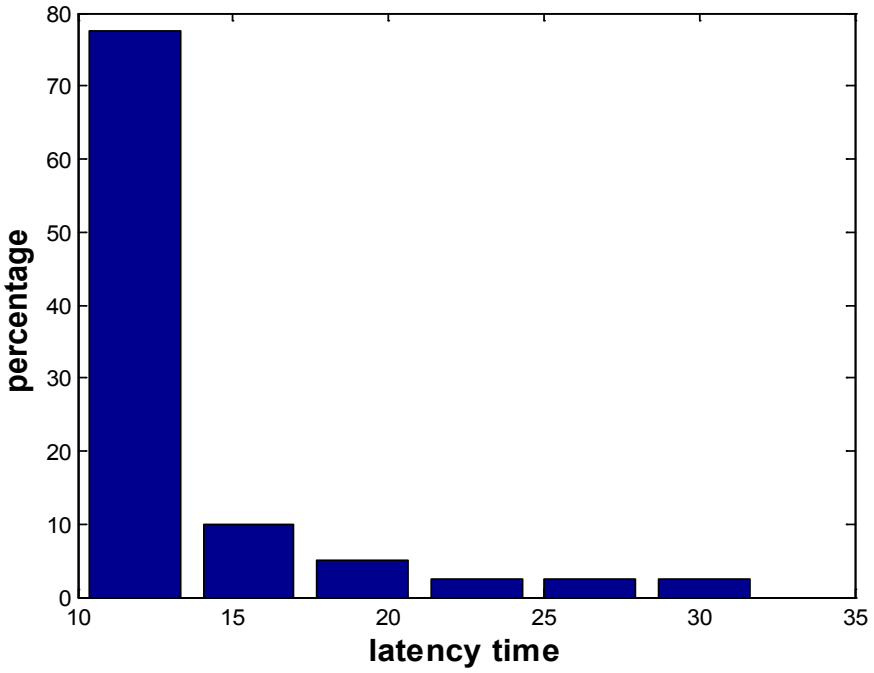


Fig. 5.11 Latency in the change point detection for 20 users.

5.6.2 Unusual Consumption Event Detection

Figs. 5.12 and 5.13 show the number of calls and call duration for user 3, where the x-axis indicates the days and y-axis indicates the number of calls and call duration (incoming, outgoing and total of them) respectively.

Fig. 5.14 shows the event days, where the x-axis indicates days and y-axis indicates the number of calls and call duration (incoming, outgoing and total of them) respectively. Fig.5.14 show there are 7 event days which are the 5th, 18th, 31st , 33rd , 58th, 62nd, and 76th, days during 106 days.

Figs. 5.15 and 5.16 show the number of calls and call duration for user 74, while Fig. 5.17 shows the event days. Fig. 5.17 show that there are 12 event days which are the 24th, 26th, 27th, 28th, 46th, 78nd, 86th, 93rd, 122nd, 130th, 131st and 153rd days during 160 days. During these event days, there is either large number of calls or longer call duration and in 6 of these days user 74 travelled to New York City and Providence, RI.

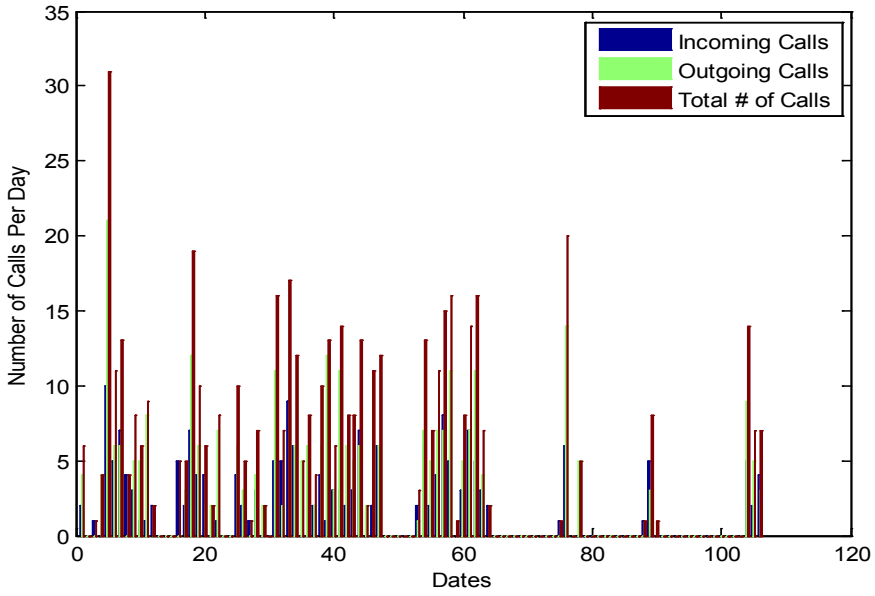


Fig. 5.12 The number of incoming, outgoing and total calls per day for user 3.

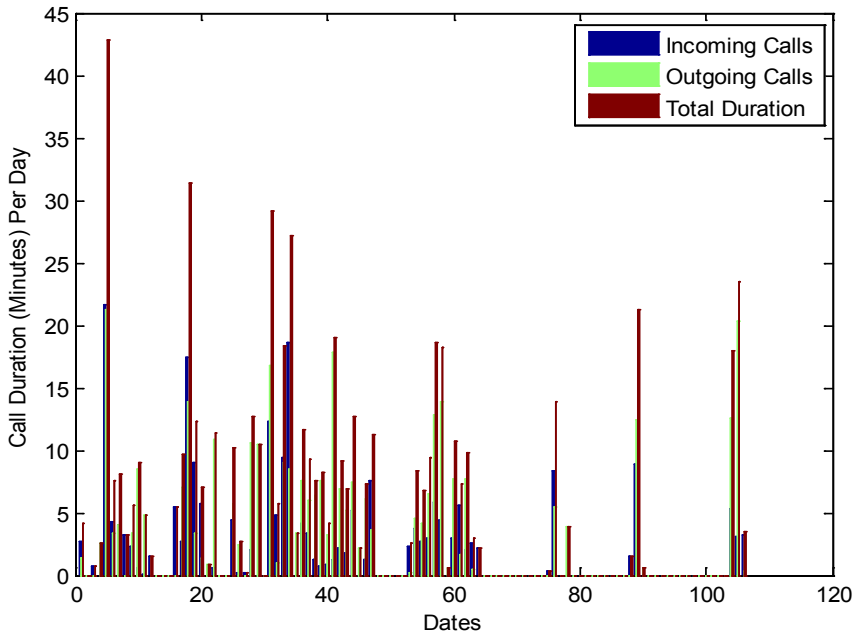


Fig. 5.13 The duration of incoming, outgoing and total calls per day for user 3.

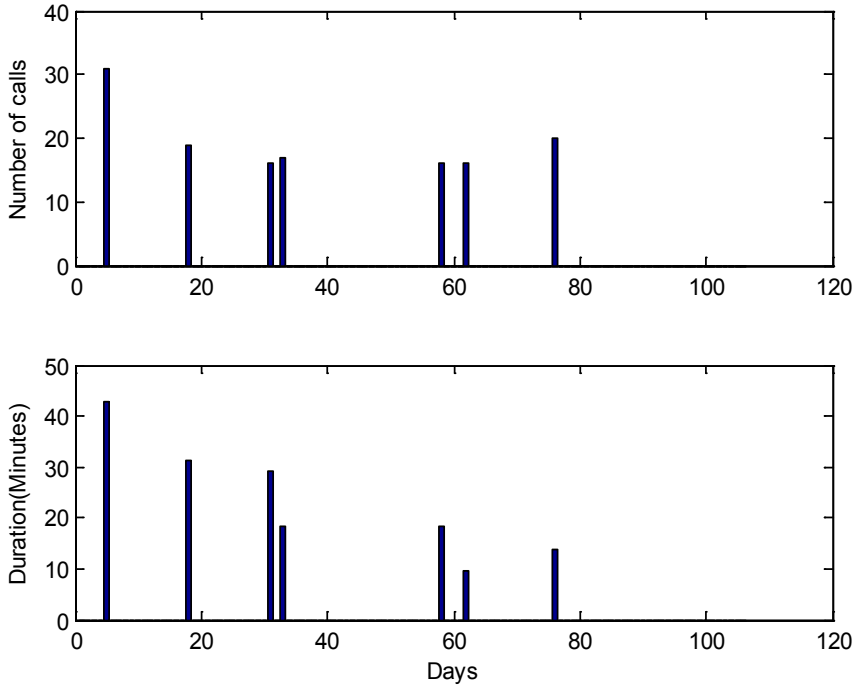


Fig. 5.14 Events for these days for user 3.

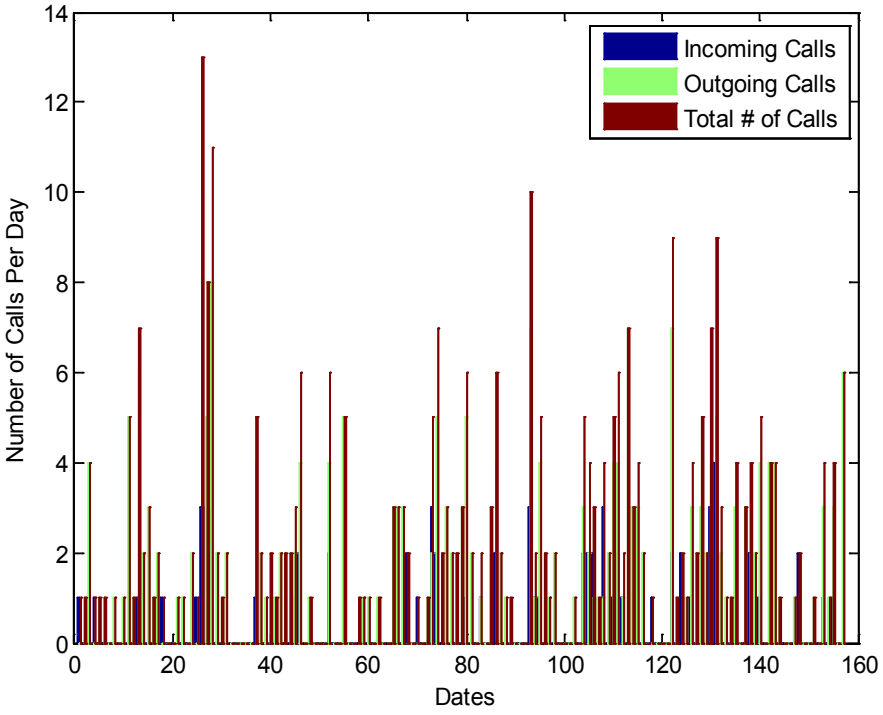


Fig. 5.15 The number of incoming, outgoing and total calls per day for user 74.

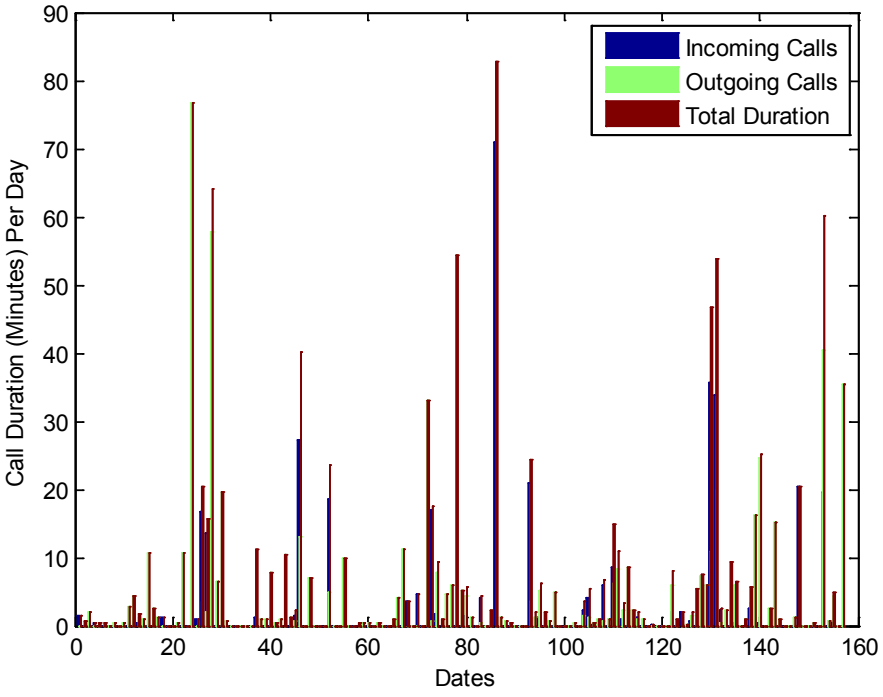


Fig. 5.16 The duration of incoming, outgoing and total calls per day for user 74.

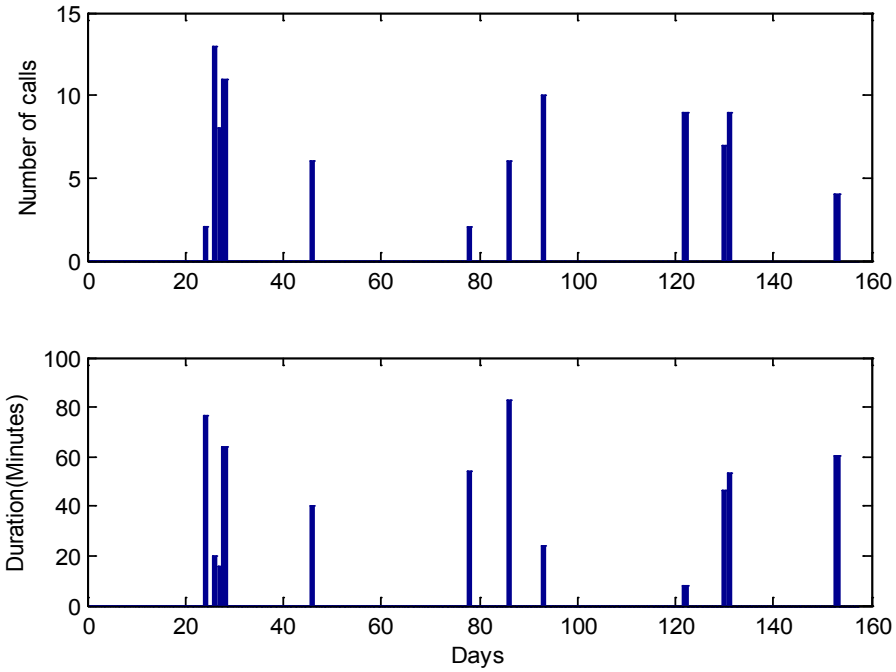


Fig. 5.17 Events for these days for user 74.

The experimental results of user3 and user74 as examples are listed in Table 5.1. In Table 5.1 the thresholds of the number of calls and duration are calculated by maximum likelihood estimates. There are two types of events, one has location change and the other has no location change.

TABLE 5.1
Event Dates and Locations

Users	Event	Days	# of contacts	# of calls	Duration (minutes)	Location	Note
User3 Threshold of # of calls per day is 15. Threshold of duration per day is 30 (minutes)	1	5th	11	31	42.9	Visit world trade center	Both large # of calls and duration
	2	18th	3	19	31.3	Visit Harvard Univ.	Both large # of calls and duration
	3	31st	8	16	29.1	Visit Harvard Univ.	Large # of calls
	4	33rd	6	17	18.3	Campus	Large # of calls
	5	58th	11	16	18	Campus	Large # of calls
	6	62nd	10	16	9.8	Campus	Large # of calls
	7	76th	7	20	13.9	Campus	Large # of calls

User74 Threshold of # of calls per day is 7 Threshold of duration per day is 36 (minutes)	1	24th	1	2	76.8	Campus	Large duration
	2	26th	4	13	20.3	Visit central square	Large # of calls
	3	27th	4	8	15.6	Campus	Large # of calls
	4	28th	2	11	64.2	At home	Large # of calls
	5	46th	1	6	40.2	Visit New York	Large duration
	6	78th	1	2	54.4	At home	Large duration
	7	86th	2	6	82.8	At home	Large duration
	9	93rd	3	10	24.5	Visit New York	Large # of calls
	10	122nd	4	9	7.9	Visit New York	Large # of calls
	11	130th	2	7	46.8	At home	Large duration
	12	131st	2	9	53.7	Home (next day visit Providence RI)	Both large # of calls and duration
	13	153rd	2	4	60	Visit New York	Large duration

For example, in Table 5.1, there are 17 calls, which are greater than 15 (the threshold of the number of calls) for user3 on the 33rd day. It is defined that there is some event in that day. For user 74 on the 24th day, although there are only 2 calls, the call duration is 76.8 minutes, which is much longer than 18 minutes (the threshold of the call duration) and there is some event in that day.

To evaluate the accuracy of the model, the actual call logs of 100 phone users are used and 20 phone users are randomly chosen. These users include students, professors and staff members. The best way to validate the results is to contact the phone users to get feedback, but because of the privacy issues it is almost impossible to do that. Thus I used hand labeling method to validate the model. I hand-labeled the events based on the number of calls, duration of calls in the day, history of call logs, location, time of arrivals, and other human intelligence factors. Table 5.2 shows the validation results, which achieved 92% accuracy.

TABLE 5.2
Validation results

Users	# of events	Threshold of # of calls per day	Threshold of duration (minutes)	Ave. # of calls per day	Ave. of duration (minutes)	False positive	False negative
3	7	15	30	5	5.5	0%	11%
14	10	9	82	4	40	0%	9%
15	9	14	21	6	7	0%	10%
16	11	8	24	4	8	0%	8%
21	8	11	56	5	18	0%	0%
22	4	22	60	11	20	0%	9%
29	12	10	14	4	7	0%	7%
33	9	6	43	1	8	0%	10%
35	5	21	76	10	21	0%	8%
38	15	15	67	10	29	0%	6%
39	13	13	52	5	14	0%	9%
50	9	16	75	7	31	0%	10%
57	8	8	15	3	5	0%	11%
72	13	12	70	6	23	0%	6%
74	13	7	36	2	7	0%	7%
78	13	10	76	3	13	0%	6%
83	12	10	28	5	9	0%	7%
85	14	9	24	4	9	0%	6%
88	11	7	16	3	4	0%	8%
95	8	4	10	2	4	0%	10%

5.7 Conclusion

In this chapter I combined wavelet denoising and the sequential change point detection method for detecting change points based on mobile phone call detail records. The data is pre-processed to reduce noise influence and the denoised data is used for the sequential change point detection to identify any new pattern in the call log data.

I also proposed the inhomogeneous Poisson process model for detecting unusual consumption events.

This work is useful for enhancing homeland security, detecting unwanted calls (e.g., spam), communication presence, and marketing among other applications. The

detection of change points in the calling pattern is an indication of the occurrence of an event in someone's life. The knowledge about certain events can be used, for example, for security purposes. Listening to all conversations of all suspects in all potential terrorist attacks is almost impossible. However, the efforts can be more concentrated if a change in behavior is observed for a certain group or individual. The experimental results show that the model can indeed detect certain events based on the observation of existing calling patterns. The use of both synthetic data as well as real data has shown promising results with regard to performance (number of detected change points) and with accuracy (number of false positives and latency).

CHAPTER 6

PATTERN RECOGNITION

6.1. Introduction

Analyzing patterns of human behavior is an area of increasing interest in a number of different applications. The automatic detection of bursts by studying patterns of human behavior is one of them and has recently attracted attention. A burst is something that happens at a given point in time and at a given place. I use *burst* to refer to a unusual large scale activity relative to normal patterns of behavior. To understand such data, I often care about both the patterns of the typical behavior, and the detection and extraction of information based on the deviations from this behavior. Almost all previous approaches for burst detection are based on text, website, and video data. The Opt-in pattern of calling patterns is similar to the bursts in the data streams.

Opt-in is an approach to e-mail or phone marketing in which customers must explicitly request to be included in an e-mail or phone call campaign or newsletter. Customers can easily choose to be removed from a mailing or phone list if the advertisements are unwanted, eliminating unsolicited emails or phone calls. People may have some interests in some advertisements for certain period of time and they do not want to receive them later. Ultimately this traffic is considered as spam and people decide to opt-out. It is very difficult for current spam filters to detect this type of traffic. Note that there can be several kinds of opt-ins and I consider opt-ins whom I show lots of interest for a short period of time and later find no interest in them but

still keep getting unwanted emails or calls from them. I call this kind of traffic as opt-in bursts.

Homeland security agents sometime want to know when the potential terrorists would like to communicate with their partners by modern telecommunication devices so that the agents can trace, scout, surveil and detect potential terroristic attacks and evidence. The Bayesian inference model is proposed to compute the willingness level of people's communications with one another at a given time. Another example of the willingness level of people's communication is on computer and telecommunication presence. The emerging of presence-aware communication allows people to quickly connect with others, whether on the road, in meetings, or working from remote locations, via the best choice of communication means. Presence awareness let users know when other people in their contact list are online. Presence information can include more user details, such as availability, location, activity, device capability and other communication preferences. Presence is used to detect and convey one's willingness and ability to talk on the phone. Presence-enabled telephony services can reduce telephone traffic, tag and improve customer satisfaction. The existing approaches provide presence for "online," "busy," "away,'' "offline," etc. The detection of opt-ins as well as the computation of willingness is presented next.

6.2. Opt-in Detection

Opt-in burst detection is related to burst detection on data streams and time series which are continuous data. However, the Opt-in behavior looks like accumulated impulses and is not continuous.

There are a lot of previous works on burst detection in data stream and time series. Kleinberg (2002) proposed an infinite-state automation model based on state transitions to detect bursts in data streams. The limitation of this model is that it is assumed that the number of documents in a document stream is uniform distribution. Fujiki et al. (2004) extended the Kleinberg's model (2002) and applied their modified model to non-uniform distribution document stream such as BBSs and blogs. Vlachos et al. (2004) used moving average technique to identify bursts. Zhu and Shasha (2003) proposed the elastic windows model and shifted wavelet tree data structure to monitor data streams and detect bursts. They applied this method to Gamma ray data set and stock exchange data set. Qin et al. (2005) defined the bursts by the ratio of aggregates of the latest two consecutive subsequences with the same length and set threshold ratio to detect burst. Chen et al. (2006) introduced the burst lasting factor and abrupt factor in the definition of bursts and used two-layered wavelet to detect them. Wang et al. (2002) used self-similarity to model bursty time series. It only needs one parameter for finding and characterizing pattern in time series. Keogh et al. (2002) used suffix tree to encode the frequency of all observed patterns and applied a Markov model to detect patterns in the symbol sequence. Fung et al. (2005) proposed parameter free probabilistic model to detect bursty events by their distribution based on time and content information. Yuan et al. (2007) proposed a burst detection algorithm based on Ratio Aggregation Pyramid and Slope Pyramid data structure to detect bursts in text streams. Vlachos et al. (2005) proposed indexing scheme, and presented interesting burst correlations to detect bursts for the financial data.

The above approaches for detecting bursts are used for text, novel, and unusual data points or segments in time-series which have either contents or traffic

data. There is no content in call detail records which is the main different from the text and website data. None of the previous work focuses on the specific problem I studied here: opt-in bursts by studying the calling pattern based on call detail records to detect opt-in bursts that reflect the human activity.

According to the features of the opt-in bursts, I proposed and investigated the dynamic size window model combined with Exponentially Weighted Moving Average (EWMA) method to detect opt-in bursts, and illustrated how to learn such model from data to both characterize normal behavior and detect anomalous opt-in bursts based on call detail records.

The opt-in bursts can be defined as sequences of densely accumulated impulses with an interval of length *w*.

Let $B=\{b_1, \ldots b_k\}$ be a subsequence of bursts contained in a sequence $S=\{s_1, \ldots, s_n\}$. The *i*th burst value is defined as

$$b_i(t) = \sum_{j=1}^{w_i} s_t \delta(t - t_j^i) \qquad (6.1)$$

where w_i is total number of impulses of the *i*th burst, i.e. the *i*th burst width, t_j^i is the occurrence point of the *j*th impulse of the *i*th burst and δ(t) is a delta function denoting the occurrence of a impulse at point $t= t_j^i$.

The *i*th burst amplitude A_i can be calculated as

$$A_i = \frac{1}{w_i} \sum_{j=1}^{w_i} s_t \delta(t - t_j^i) \qquad (6.2)$$

where s_t is the value of an pulse at point t.

To detect the bursts, we define the sliding window SW_k as

$$SW_k(t) = A_k rect(\frac{t - t_m}{\tau_k}) \quad (6.3)$$

where A_k is the amplitude of a sliding window k, rect((t-t_m)/τ_k) is rectangle function denoting the occurrence point of a burst at time t=t_m and τ_k is the width of a sliding window k.

Definition of an opt-in burst

A collection of call log data can be represented as

$$C = < (t_1, a_1, d_1, l_1), (t_2, a_2, d_2, l_2), ... (t_n, a_n, d_n, l_n) >,$$

where t_i is a time point, d_i is a call duration, l_i is a location and a_i is a pair of actors, caller-callee $< s_i, r_i >$ where s_i is an actor who initiates a call at time t_i and r_i is an actor who receive a call. An opt-in burst is defined as a subset $E \subset C$ of a tuple

$$E = \{(t_1, a_1, d_1, l_1), (t_2, a_2, d_2, l_2), ... (t_m, a_m, d_m, l_m)\}$$

such that $0 < count(d_i) < N_{thres}$ in the time period $\Delta t = t_m - t_1$, where N_{thres} is threshold which can be estimated from the historical data.

The sequence S is processed by exponentially weighted moving average (EWMA), and then the dynamic size sliding windows are applied to detect opt-in bursts. The EWMA places more emphasis on the most recent data. Therefore it would be more useful in dynamic systems.

Let $S = \{s_1, ..., s_n\}$ be a sequence. Then the moving average (*MA*) is given by

$$\bar{s}_k = \frac{1}{M} \sum_{i=k-M+1}^{k} s_i \quad (6.4)$$

where $\bar{s}_k$ is moving average of the k's instance and M is the number of latest values.

The EWMA can be derived from *MA* as

$$\bar{s}_k = \alpha \bar{s}_{k-1} + (1-\alpha) s_k$$

where $0 \leq \alpha < 1$ is a constant. This is a recursion function. By recursion we may have

$$\bar{s}_k = (1-\alpha)s_k + \alpha(1-\alpha)s_{k-1} + \alpha^2(1-\alpha)s_{k-2} + \alpha^3 \bar{s}_{k-3} \qquad (6.5)$$

where $0 \leq \alpha < 1$ is a constant.

If I keep on expanding $\bar{s}$ terms on the right hand side, the contribution of older values of s_k is weighted by increasing powers of α. Since α is less than 1, the contribution of older values of s_k becomes progressively smaller.

6.3 Willingness Level Inference

When a caller wants to make a call, he/she would like to know if the callee is in a mood to receive a call. In other words the caller would like to know when it is a good time to call the particular callee. From the network traffic control point of view, this can reduce traffic congestion since the caller knows the callee's willingness level, so the caller might not initiate a call and this also saves the caller's available minutes. The chance is estimated based on the time of the day, call duration, and the location. The Bayesian inference model is proposed to compute the willingness level of a receiver at a given time. As one of its applications, the Willingness Calculator (WC) is proposed for computing the willingness level of the callee, which can be deployed at the callee's Home Location Register (HLR in a cellular network). The WC service flow diagram is shown in Fig. 6.1.

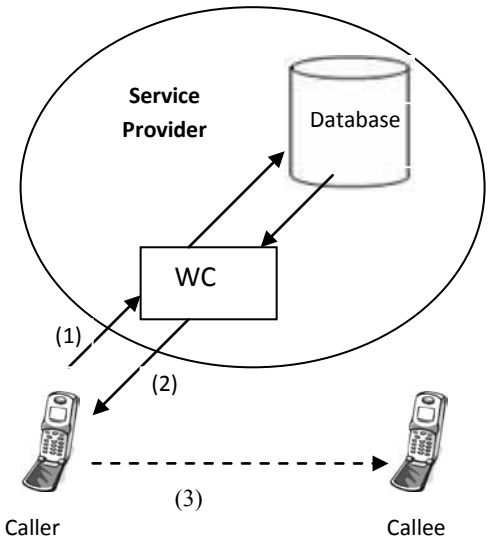


Fig. 6.1 Basic service flow diagram.

Let X and Y be two events. By conditional probability rule [Nilsson 1998], the probability of an event X given Y is

$$P(X \mid Y) = \frac{P(X,Y)}{P(Y)}$$

where $P(X,Y)$ is the joint probability.

By the chain rule of conditional probability (Nilsson 1998), we have

$$P(X,Y) = P(X \mid Y)P(Y) \tag{6.6}$$

Since the order of X and Y does not matter in Eq. (20), we have

$$P(Y, X) = P(Y \mid X)P(X)$$

Since $P(X,Y) = P(Y,X)$, we have

$$P(X \mid Y)P(Y) = P(Y \mid X)P(X)$$

Thus, we have Bayes' theorem:

$$P(X \mid Y) = \frac{P(Y \mid X)P(X)}{P(Y)} \tag{6.7}$$

In (6.7), $P(X \mid Y)$ is called posterior probability, $P(Y \mid X)$ is referred to as likelihood and $P(X)$ is prior probability.

Let X = (incoming call = accept, incoming call = missed).

Let Y = (T_i, D_j, Loc_l), where T_i is time interval, i = 0, 1, 2, …23, (e.g. : 0 – 1 O'clock), D_j is a day, j=1,2, …, 7 i.e. D_1=Sunday, D_2=Monday, … D_7= Saturday, Loc_l = location name, l=1, 2, …n .

Then by Bayes theorem the willingness level to accept a call is

$$P(inco\min g = accept \mid T_i, D_{j,} Loc_l) = \frac{P(T_i, D_j, Loc_l \mid inco\min g = accept)P(inco\min g = accept)}{P(T_i, D_j, Loc_l)}$$

(6.8)

6.4 Experimental Results and Discussions

6.4.1 Opt-in Detection

Fig. 6.2 shows the opt-in bursts for user1 with his/her communication partner123, where the x-axis indicates the days and the y-axis indicates the number of calls and the call durations including incoming and outgoing calls respectively. The user1 first called his/her communication partner123. Fig. 6.3 shows the opt-in bursts for user2 with his/her communication partner17. For 20 phone users with 1863 communication partners, I found that 1.4% of them were opt-in bursts.

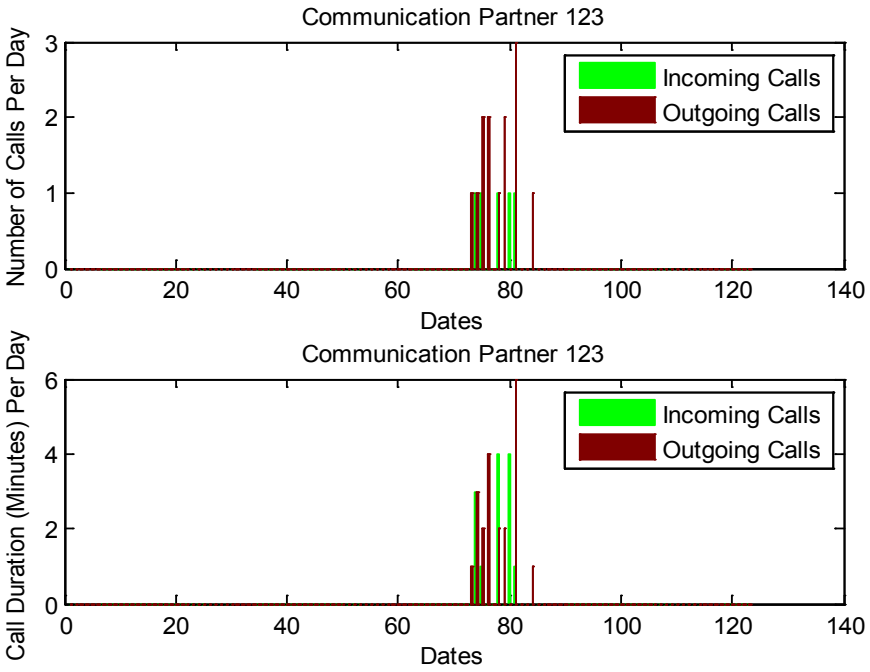


Fig. 6.2 The opt-in bursts for user 1.

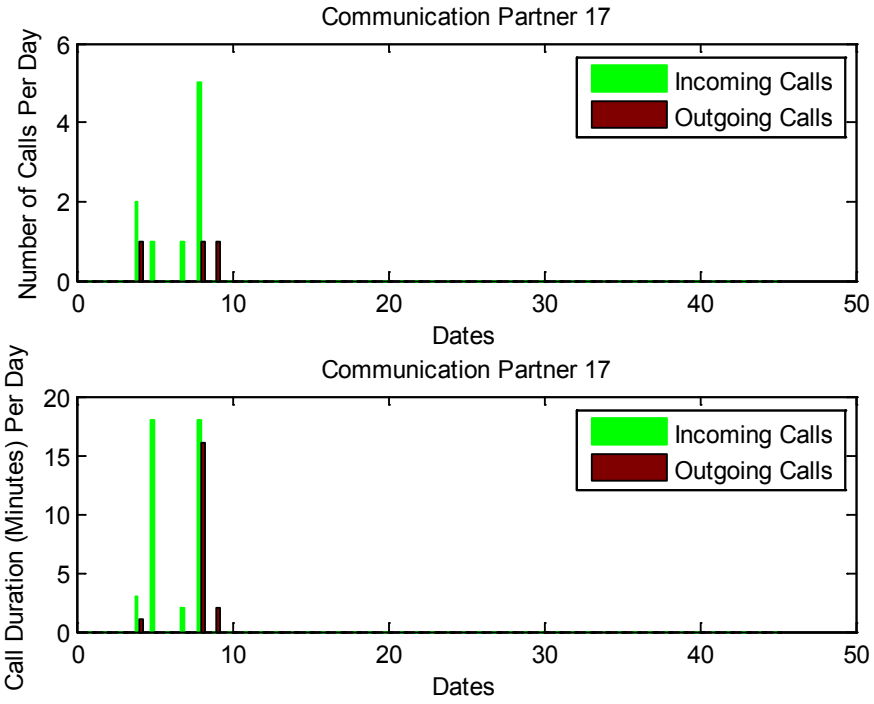


Fig. 6.3 The opt-in bursts for user 2.

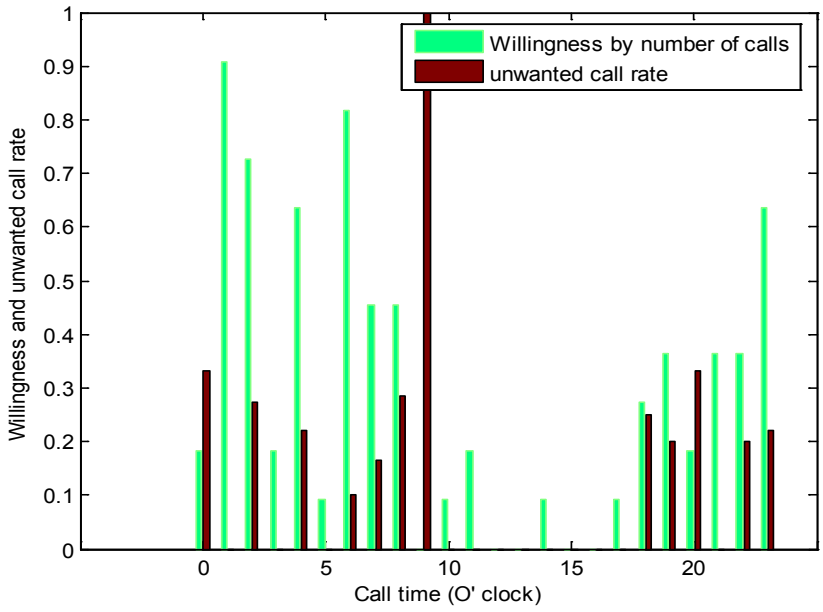
6.4.2 Willingness Level Inference

The willingness level to receive calls and the corresponding unwanted call rate for users were calculated for one hour interval from 0 – 23 o'clock from Sunday to Saturday.

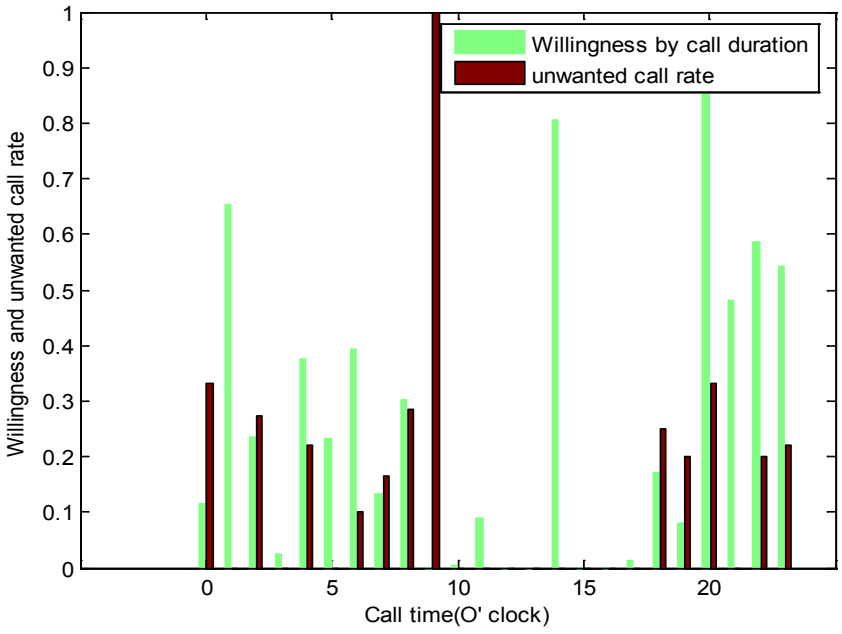
In the Figs. 6.4 and 6.5 the x-axis indicates the calling time for incoming as well as outgoing calls for 24 hours on Sunday and Monday and the y-axis indicates the willingness level for a second year graduate student user. In this graph, missed calls were considered as unwanted calls and these were checked with willingness level. When the willingness level is low, then there are more missed calls.

Fig. 6.4(b) describes the willingness calculated based on total talk time. Fig. 6.4 (a) and (b) show that when the user is more willing to receive calls, the number of missed calls decreases. The missing calls of the receiver means these are unwanted calls at a given time. For example, in Fig. 6.4 (a) the willingness level is 0.7 (70%) and corresponding unwanted call rate is 0.28 (28%) between 2 to 3 o'clock. Consistently, the willingness level is 0.2 (20%) and the corresponding unwanted call rate is 0.33 (33%) between 0 to 1 o'clock.

Fig. 6.5 shows the willingness level of this user on Mondays. Fig. 6.5 (a) and (b) show that higher willingness level corresponds to lower unwanted call rate. For example, in Fig. 6.5 (a) the willingness level is 0.84 (84%) and the corresponding to unwanted call rate is 0.1 (10%) between 0 to 1 o'clock. Similarly the willingness level is 0.1 (10%) and the corresponding unwanted call rate is 0.5 (50%) between 23 to 0 o'clock.

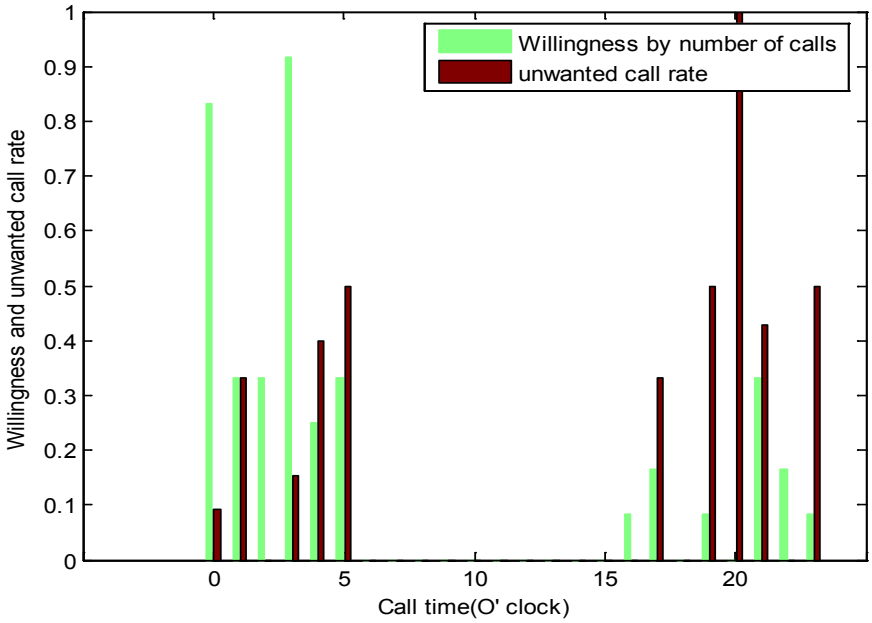


(a)

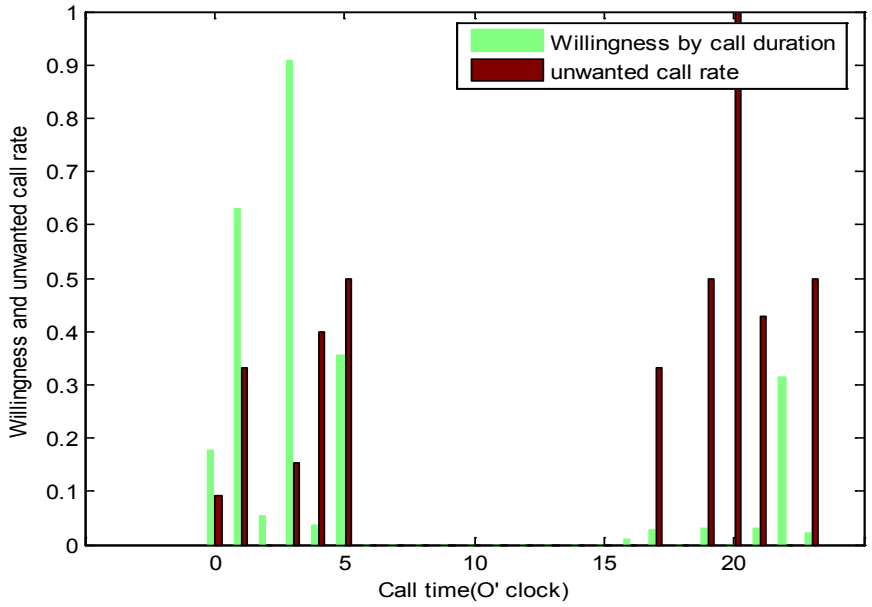


(b)

Fig. 6.4 Willingness level and unwanted rate on Sundays. (a) Willingness level compared to unwanted call rate (rejected/missed) (b) computed willingness level based on talk time.



(a)



(b)

Fig. 6.5 Willingness level and unwanted rate on Monday. (a) Willingness level and unwanted rate by number of calls (b) Willingness level and unwanted rate by call duration.

Fig. 6.6 shows the willingness level of this second year graduate student user based on the number of calls he received from Sunday to Saturday. Here *x*-axis represents time of the day, and the *y*-axis represents 7 days of a week. The first unit on the *y*-axis represents Sunday and the last unit represents Saturday.

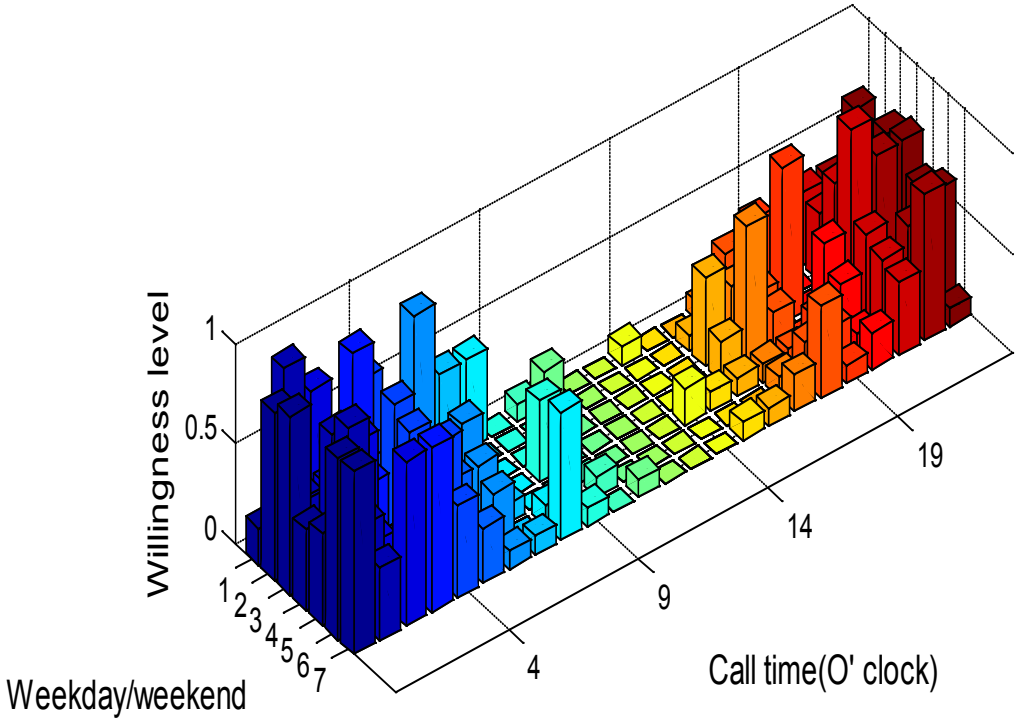


Fig. 6.6 Willingness level during 24 hours from Sunday to Saturday.

I validated the willingness with respect to number of miss or rejected calls. The accuracy was measured by the unwanted call rate over the range of different willingness levels. The unwanted call rate is a ratio of number of miss calls to the total number of calls at given time period. The assumption is that a miss call is an unwanted call.

Table 6.1 and Fig. 6.7 describe the experimental results for 10 phone users. In Table 6.1 the results show that the model achieves good performance. For example,

when the willingness level is 0 - 30%, the average unwanted rate is 43.32% with standard error 1.79%. Whereas the mean unwanted rate is 4.84% for the willingness of 71-100%. The higher the willingness level, the lower the unwanted rate, and vice versa.

Table 6.1 Unwanted Call Rate Corresponding to the Willingness Level

Phone users	Number of incoming calls	Number of unwanted calls	Unwanted rate (%)		
			Willingness level (%)		
			0-30	31-70	71-100
1 (student)	564	128	41.3	14.3	7.4
2 (staff)	230	68	45.7	8.1	2.7
3 (professor)	341	52	32.7	11.5	3.3
4 (student)	563	88	45.9	11.1	6.3
5 (student)	1007	195	35.1	17.8	8.8
6 (professor)	255	53	42.4	13.8	1.1
7 (staff)	186	55	47.6	14.6	2.1
8 (student)	487	180	49.8	16.9	4.6
9 (student)	361	143	48.9	12.0	4.9
10 (student)	286	69	43.8	10.1	7.2

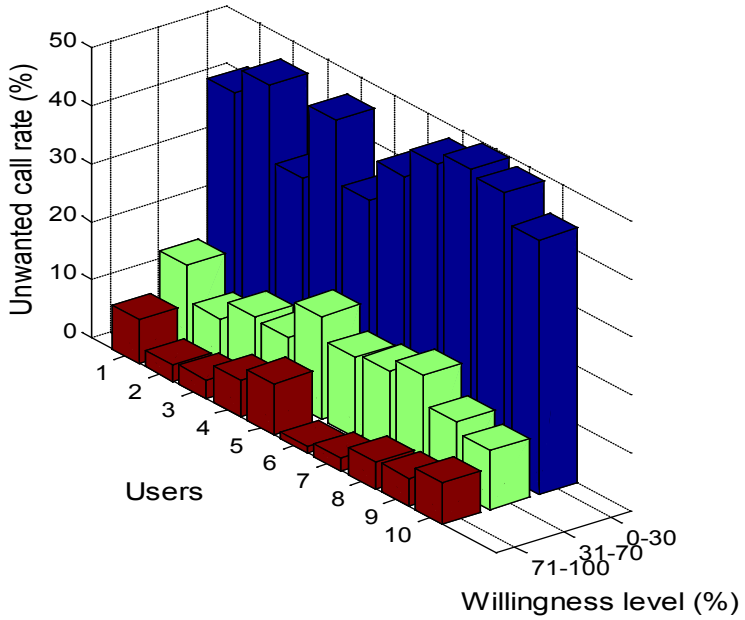


Fig. 6.7 Willingness level (%) vs. unwanted call rate (%) for 10 users.

6.5 Conclusion

In this chapter I proposed the approach combining the dynamic sliding windows model with exponentially weighted moving average for detecting opt-in bursts.

I also proposed the Bayesian inference model to quantify the willingness level to receive a call at a given time.

This work is useful for detecting unwanted calls (e.g., spam), marketing, etc.

CHAPTER 7

CONCLUSIONS

In this dissertation I proposed a socioscope model for social network and human behavior analysis based on mobile phone call detail records. The socioscope model consists of several components including data extraction and transformation, network visualization, zoom, scale, and analysis tools which are used for analyzing network structures, discovering and quantifying social groups and events, quantifying and predicting relationships. It is extensible and new tools can be added as additional features are identified. Because of the diversity and complexities of human social behavior, one technique cannot detect different features of human social behavior. I used multiple probability and statistical methods to analyze human social behavior. First I proposed a new reciprocity index to measure the level of reciprocity between users and their communication partners. Then I integrated the new reciprocity index into the proposed affinity model for quantifying interpersonal relationships and social group identification. Since interpersonal relationships dynamically change over time, I mapped call-log data into time series of social tie strengths by my affinity model so that social tie strengths are functions of time. Furthermore, I applied seasonal auto regressive integrated moving average (SARIMA) model to predict social tie strengths.

To investigate human behavior changes over time, I combined wavelet denoising and sequential detection methods for detecting change points. Change point detection methods do not deal with bursts of short width in time series. In order to overcome this shortcoming of change point detection methods, I proposed the inhomogeneous Poisson model for detecting unusual consumption events.

In order to investigate opt-in behavior which is a new emerging phenomenon of phone calls, I proposed the dynamic size window model combined with exponentially weighted moving average (EWMA) method for detecting opt-in bursts. When we are in busy time such as during teaching or taking classes, we do not like to take phone calls and the calls will be forwarded to voice box. So we have some particular time when we cannot take phone calls. In order to know when it is a good time to call somebody, I proposed a Bayesian inference model for quantifying willingness level to communicate with each other based on human telecommunication patterns. We may quantify relationships for a short-term period, say a month, or a long-term period, say a year or more, using this approach by adjusting the parameters. Errors appeared when the number of calls is very few. However, these kinds of cases seldom happened in the experiments.

The experimental results show that my approaches are effective.

This work is useful for homeland security, the detection of unwanted calls(e.g., spam), telecommunication presence and marketing, etc. In future work I plan to analyze and study the social network dynamics and evolution.

REFERENCES

Backstrom, L., Huttenlocher, D. and Kleinberg, J. 2006. "Group formation in large social networks: membership, growth, and evolution." In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 44-554.

Baron, M. 2000. "Nonparametric adaptive change-point estimation and on-line detection." *Sequential Analysis*, 19: 1–23.

Baker, F. B. and Hubert, L. J. 1981. "The analysis of social interaction data." *Social Methods Res.* 9: 339–361.

Barabási, A. L., Albert, R., Jeong, H. and Bianconi, G., 2000, "Power-Law Distribution of the World Wide Web." *Science* 287: 2115.

Barabási, A. L., Jeong, H., Neda, Z., Ravasz, E., Schubert, A. and Vicsek, T., 2002, "Evolution of the social network of scientific collaborations." *Physica* A 311: 590-614.

Baron, M and Granott, N. 2003. "Consistent estimation of early and frequent change points." In Y. Haitovsky, H. R. Lerche, Y. Ritov, eds., Foundations of Statistical Inference, In *Proceedings of the Shoresh Conference* 2000, Springer, 181–192.

Beauchamp, 1965. An improved index of centrality. *Behavioral Science*, 10: 161-163.

Borgs, C., Chayes, J., Mahdian M. and Saberi, A. 2004. "Exploring the community structure of newsgroups." In *Proceeding of 10th ACM International. Conference on Knowledge Discovery and Data Mining*.

Box, G. E. P., Jenkins, G. M. and Reinsel, G. C. 1994. *Time series analysis: Forecasting and control*, 3rd ed., Prentice-Hall, Englewood Cliffs, NJ.

Brants, T. and Chen, F. 2003. "A System for New Event Detection." In *Proceedings of international ACM SIGIR conference*: 330-337.

Brass, D. J. 1995. "A social network perspective on human resources management." *Research in Personnel and Human Resources Management*, 13: 39–79.

Carley, K. M., Lee, J. and Krackhardt, D. 2002. "Destabilizing networks." *Connections*, 24(3): 79–92.

Bregni, S., Cioffi, R. and Decina, M. 2006. "An Empirical Study on Statistical Properties of GSM Telephone Call Arrivals." In *Proceeding of IEEE Global Telecommunications Conference*.

Brickley, D. and Miller, L. 2004. "FOAF Vocabulary Specification." Namespace Document, http://xmlns.com/foaf/0.1

Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A. and Wiener, J. 2000. "Graph structure in the web." *Computer Networks*, 33 (2): 309–320.

Candia, J., Gonzalezm, M. C., Wang, P., Schoenharl, T., Madey, G., Barabasi, A. 2008. "Uncovering individual and collective human dynamics from mobile phone records." *Journal of Physics A: Mathematical and Theoretical*. 41: 224015.

Cangussu, J. W. and Baron, M. 2006. "Automatic identification of change points for the system testing process." In *Proceedings of the 30th Annual IEEE International Computer Software and Applications Conference.* Chicago, IL, Sept. 18-21.

Carley, K. 2003. "Dynamic Network Analysis." Forthcoming in in the Summary of the NRC workshop on Social Network Modeling and Analysis, Breiger, R. and Carley, K. M. (eds), National Research Council, http://stiet.si.umich.edu/researchseminar/Winter%202003/DNA.pdf.

Carlin, B. P., Gelfand. A. E. and Smith, A. F. M. 1992. "Hierarchical Bayesian analysis of change point problems." *Applied Statistics* 41: 389-405.

Carrington, P. J., Scott, J. and Wasserman, S. 2005, *Models and Methods in Social Network Analysis*, Cambridge University Press.

Chen, J. and Gupta, A. K. 1997. "Testing and locating change points with application to stock prices." *Journal of the American Statistical Association*. 92: 739–747.

Chen, F., Farahat, A. and Brants, T. 2003. "Story Link Detection and New Event Detection are Asymmetric." In *Proceedings of the Human Language Technology Conference*.

Chen, T., Wang, Y., Fang, B. and Zheng, J. 2006. "Detecting Lasting and Abrupt Bursts in Data Streams Using Two-Layered Wavelet Tree." In *Proceedings of the Advanced International Conference on Telecommunications*.

Chi, E. H. and Mytkowicz, T. 2008. "Understanding the efficiency of social tagging systems using information theory." In *Proceedings of the nineteenth ACM conference on Hypertext and hypermedia*, 81–88.

Chi, Y., Song, X., Zhou, D., Hino, K. and Tseng. B. L. 2007. "Evolutionary spectral clustering by incorporating temporal smoothness." In *Proceedings of the Thirteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 153–162.

Chu, P. and Zhao, X. 2004. "Bayesian Change-Point Analysis of Tropical Cyclone Activity: The Central North Pacific Case." *Journal of Climate*, 17: 4893-4901.

Clauset, A., Newman, M. E. J. and Moore, C. 2004. "Finding community structure in very large networks." *Physical Review E*, 70: 066111.

Coffman, T., Greenblatt, S. and Marcus, S. 2004, "Graph-Based Technologies for Intelligence Analysis." *Communications of the ACM*, 47(3): 45-47.

Crucitti, P., Latora, V., and Porta, S. 2006. "Centrality measures in spatial networks of urban streets." *Physical Review,* E 73: 036125.

Dasgupta, K., Singh, R., Viswanathan, B., Chakraborty, D., Mukherjea, S. and Nanavati, A. 2008. "Social Ties and their Relevance to Churn in Mobile Telecom Networks." In *Proceedings of the 11th ACM international conference on extending database technology: Advances in database technology*.

Desikan, P. and Srivastava, J. 2004, "Mining temporally evolving graphs." In *Proceedings of the sixth WEBKDD workshop in conjunction with the 10th ACM SIGKDD conference*, August 22.

Dill, S., Kumar, R., McCurley, K., Rajagopalan, S., Sivakumar, D. and Tomkins, A. 2001. "Self-similarity in the Web." In *Proceedings of the 27th International Conference on Very Large Data Bases*.

Donath, J., Karahalios, and Viegas, K. F. 1999. "Visualizing conversation." In *Proceeding of Hawaii International Conference on System Sciences*. 32, Jan.

Donoho, D. and Johnstone, I. 1993. "Adapting to Unknown Smoothing via Wavelet Shrinkage." *J. Amer. Statist. Association*, 90: 1200-1224.

Donoho, D. and Johnstone, I. 1994. "Ideal Spatial Adaptation by Wavelet Shrinkage." *Biometrica*, 81: 425-455.

Doreian, P., Batageli, V. and Ferligoj, A. 2005. *Generalized Blockmodeling*, M. Granovetter, Ed. Cambridge, UK: Cambridge University Press.

Dorogovtsev, S.N., and Mendes, J. F. F., 2003. *Evolution of Networks: From Biological Nets to the Internet and WWW*. Oxford University Press, New York.

Du, N., Faloutsos, C., Wang, B. and Akoglu, L. 2009. "Large human communication networks: patterns and a utility-driven generator." In *Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining*.

Eagle, N. 2008. "Behavioral inference across cultures: using telephones as a cultural lens." *IEEE Intelligent Systems* 23(4): 62-64.

Eagle, N., Pentland, A and Lazer, D. 2009. "Inferring Social Network Structure using Mobile Phone Data." In *Proceedings of the National Academy of Sciences*. 106(36): 15274-15278

Erdman, C. and Emerson, J. W. 2008. **"**A fast Bayesian change point analysis for the segmentation of micro-array data**."** *Bioinformatics*. 24(19): 2143-2148.

Fannes, M. and Spincemaile, P. 2003. The mutual affinity of random measures. *Periodica Mathematica Hungarica*, 47: 51–71.

Farrell, S., Campbell, C. and Myagmar, S. 2005, "Relescope: An Experiment in Accelerating Relationships." In *Proceedings of Conference on Human Factors in Computing Systems*, April 2005, Portland, 2-7.

Fearnhead, P. 2006. "Exact and efficient Bayesian inference for multiple Change point problems." *Stat Comput* 16: 203–213.

Fellman, P. V. and Wright, R. 2004, "Modelling Terrorist Networks - Complex Systems at the Mid-Range." http://www.psych.lse.ac.uk/complexity/Conference/FellmanWright.pdf.

Fisher, D. 2005. "Using egocentric networks to understand communication." *IEEE Internet Computing,* 9(5): 20–28.

Flake, G., Lawrence, S. and Lee, G. 2000. "Efficient Identification of Web Communities." In *Proceedings of the sixth ACM SIGKDD international conference on knowledge discovery and data mining.150-160.*

Flake, G., Lawrence, S., Giles, C. L. and Coetzee, F. 2002. "Self-Organization and Identification of Web Communities." *IEEE Computer*, March, 35:3.

Flake, G. W., Tarjan, R. E. and Tsioutsiouliklis, K. 2004. "Graph Clustering and Minimum Cut Trees." *Internet Math*. 1.

Floría, L. M., Gracia-Lázaro, C., Gómez-Gardeñes, J and Moreno, Y. 2009. "Social network reciprocity as a phase transition in evolutionary cooperation." *Phys. Rev*. E **79**: 026106.

Frank, O. 2002, "Using centrality modeling in network surveys." *Social Networks*, 24: 385.

Freeman, L. C. 1979, "Centrality in social networks: Conceptual clarification." *Social Networks*, 1: 215.

Frivolt, G., and Bielikov, M. 2005. "An approach for community cutting." In *Proceedings of the 1st International Workshop on Representation and Analysis of Web Space,* 49–54.

Fujiki, T., T, Nanno, Suzuki, Y. and Okumura, M. 2004. "Identification of Bursts in a Document Stream." In *Proceedings of First International Workshop on Knowledge Discovery in Data Streams*.

Fung, G. P. C., Yu, J. X.; Yu, P. S. and Lu, H. 2005. "Parameter free bursty events detection in text streams." In *Proceedings of the 31st international conference on very large data bases*.

Gabor, D. 1946. "Theory of Communication." *Journal of the IEE* 93: 429-457.

Garlaschelli, D. and Loffredo, M. I. 2004. "Patterns of link reciprocity in directed networks." *Phys. Rev. Lett.* 93: 268701.

Gilbert, E. and Karahalios, K. 2009. "Predicting Tie Strength with Social Media." In *Proceedings of the 27th international conference on Human factors in computing systems*. 211-220.

Girvan, M. and Newman, M. E. J. 2002. "Community structure in social and biological networks." In *Proceedings of the National Academy of Sciences of United States of America* 99 (12): 7821–7826.

Gooijer, J and Hyndman, R. J. 2006. "25 years of time series forecasting." *International Journal of Forecasting* 22 (3): 442- 473.

Gouldner A. W. 1960. "The norm of reciprocity: A preliminary statement." *American Sociological Review* 25: 161-178.

Gulati, R., 1998. "Alliances and networks." *Strategic Management Journal* 19: 293-317.

Guralnik, V. and Srivastava, J. 1999. "Event detection from time series data." In *Proceedings of the fifth ACM SIGKDD International Conference*, 33–42.

Harris, J. W. and Stocker, H. 1998. Maximum Likelihood Method. §21.10.4 in *Handbook of Mathematics and Computational Science*. New York: Springer-Verlag, 824.

Hawkins, D. M. 1977. "Testing a sequence of observations for a shift in Location." *Journal of the American Statistical Association* 72: 180-186.

He, Q., Chang, K. and Lim, E. P. 2006a. "A model for anticipatory event detection." In *Proceedings of the 25th International Conference on Conceptual Modeling* (ER), Springer LNCS 4215: 168-181.

He, Q., Chang, K. and Lim, E. P. 2006b. "Anticipatory event detection via sentence classification." In *Proceedings of IEEE International Conference on Systems, Man, and Cybernetics*, 1143-1148.

Hidalgo, A. C. and Rodriguez-Sickert, C. 2008. "The Dynamics of a Mobile Phone Network." *Physica* A 387: 3017-3024.

Hogan, B. and Fisher, D. 2006. "Email reciprocity and personal communication bias." In *Proceeding of NetSci2006*, May 22–25, Bloomington, IN, USA.

Holme, P. 2003, "Network dynamics of ongoing social relationships." *Europhys. Lett*. 64: 427-433.

Holme, P., Edling, C. and Liljeros, F. 2004. "Structure and Time Evolution of an Internet Dating Community." *Social Networks* 26: 155-174.

Holme, P. and Newman, M. 2006. "Nonequilibrium phase transition in the coevolution of networks and opinions." *arXiv physics*/0603023, March.

Hopcroft, H., Khan, O., Kulis, B. and Selman, B. 2003. "Natural communities in large linked networks." In *Proceeding of 9th International Conference on Knowledge Discovery and Data Mining*.

Huang, Z., Li, X. and Chen, H. 2005. "Link Prediction Approach to Collaborative Filtering." In *Proceedings of the 5th ACM/IEEE-CS joint conference on digital libraries*, June 2005, 7-11.

Ihler, A., Hutchins, J. and Smyth, P. 2006. "Adaptive Event Detection with Time–Varying Poisson Processes." In *Proceedings of the ACM SIGKDD International Conference*. 207-216.

Jackson, M.O., and Rogers, B. W. 2007. "Meeting Strangers and Friends of Friends: How Random are Socially Generated Networks?" *American Economic Review*, 97: 890-915.

Jawerth, B. and Sweldens, W. 1994. "An Overview of Wavelet Based Multiresolution Analysis." *SIAM Review* 36: 377-412.

Johnson, T. D., Elashoff, R. M. and Harkema, S. J. 2003. "A Bayesian change-point analysis of electromyographic data: detecting muscle activation patterns and associated applications." *Biostatistics*. 4:143– 164.

Johnstone, I and Silverman, B. W. 1997. "Wavelet Thresholding Estimators for Data with Correlated Noise." *Journal of Roy. Stat. Soc*. 59 (B): 319-351.

Kahanda, I and Neville, J. 2009. "Using Transactional Information to Predict Link Strength in Online Social Networks." In *Proceedings of the 3rd International AAAI Conference on Weblogs and Social Media*, 74-81.

Katz, L. and Powell, J. 1955. "Measurement of the tendency toward reciprocation of choice." *Sociometry* 18: 659-665.

Katz, L. and Wilson, T. 1956. "The variance of the number of mutual choices in sociometry." *Psychometrika* 21: 299-304.

Keogh, E., Lonardi, S. and Chiu, B. 2002. "Finding surprising patterns in a time series database in linear time and space." In *Proceedings of the eighth ACM SIGKDD International Conference*, 550–556.

Kleinberg, J. 2002. “Bursty and hierarchical structure in streams.” In *Proceedings of the eighth ACM SIGKDD International Conference*, 91–101.

Koka, R. B., Madhavan, R., and Prescott, J. E., 2006. “The Evolution of Inter-firm Networks: environmental effects on patterns of network change.” *Academy of Management Review* 31 (3): 721-737.

Kossinets, G. and Watts, D. 2006. “Empirical analysis of an evolving social network.” *Science*, 311:88 – 90.

Krackhardt, D. 1988. “Predicting With Networks - Nonparametric Multiple-Regression Analysis of Dyadic Data.” *Social Networks* 10 (4): 359-381.

Kumar, R., Novak, J., Raghavan, P. and Tomkins, A. 2003. “On the bursty evolution of blogspace.” In *Proceedings of the International Conference on World Wide Web*, 568–576.

Kumar, R., Novak, J., Raghavan, P. and Tomkins, A. 2004. “Structure and Evolution of blogspace.” *Communications of ACM*, 47(12): 35-39.

Kumar, R., Novak, J. and Tomkins, A. 2006. “Structure and Evolution of on line social networks.” In *Proceedings of the twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.*

Kumaran, G. and Allan, J. 2004. ”Text Classification and Named Entities for New Event Detection.” In *Proceedings of international ACM SIGIR Conference*, 297-304.

Kurdia, A., Daescu, O., Ammann, L., Kakhniashvili, D., and Goodman, S. 2007. “Centrality measures for the human red blood cell interactome.” *Engineering in Medicine and Biology Workshop, IEEE Dallas*, 98–101.

Kurucz, M., Benczur, A., Csalogany, K. and Lukacs, L. 2007. "Spectral Clustering in Telephone Call Graphs." In *Proceedings of the Joint 9th WEBKDD and 1st SNA-KDD Workshop*.

Leskovec, J., Kleinberg, J. and Faloutsos, C. 2005. "Graphs over time: densification laws, shrinking diameters and possible explanations." In *Proceedings of the eleventh ACM SIGKDD international conference on knowledge discovery and data mining*, New York, NY, 177–187.

Leskovec, J., Backstrom, L., Kumar, R. and Tomkins, A. 2008. "Microscopic evolution of social networks." In *Proceeding of the 14th ACM SIGKDD international conference on knowledge discovery and data mining*.

Li, X. and Croft, B. W. 2005. "Novelty Detection Based on Sentence Level Patterns." In *Proceedings of ACM CIKM*, 744-751.

Li, Z., Wang, B. and Li, M. et al. 2005. "A Probabilistic Model for Retrospective News Event Detection." In *Proceedings of international SIGIR conference*, 106-113.

Liben-Nowell, D. and Kleinberg, J. 2007. "The link prediction problem for social networks." *Journal of the American Society for Information Science and Technology* 58 (7) 1019-1031.

Lin, Y., Sundaram, H., Chi, Y., Tatemura, J. and Tseng, B. 2006. "Discovery of Blog Communities based on Mutual Awareness." In *Proceedings of the 3rd Annual Workshop on Weblogging Ecosystem: Aggregation, Analysis and Dynamics, 15th World Wide Web Conference*.

Lin, Y., Sundaram, H., Chi, Y., Tatemura, J. and Tseng, B. 2007. "Blog Community Discovery and Evolution Based on Mutual Awareness Expansion." In *Proceedings of* 2007 *IEEE/WIC/ACM International Conference on Web Intelligence*, 48-56.

Lund R. and Reeves, J. 2002. "Detection of undocumented change points: A revision of the two-phase regression model." *Journal of Climate* 15: 2547–2554.

Luo, G., Tang, C. and Yu, P. 2007. "Resource-Adaptive Real-Time New Event Detection." In *Proceedings of International ACM SIGMOD Conference on Management of Data*.

Ma, H.-W., and Zeng, A.-P. 2003. "The connectivity structure, giant strong component and centrality of metabolic networks." *Bioinformatics* 19 (11): 1423–1430.

Marsden, P. V. 2002, "Egocentric and Sociocentric Measures of Network Centrality." *Social Network* 24: 407.

Memon, N., Larsen, H. L., Hicks, D. L., and Harkiolakis, N. 2008. "Detecting hidden hierarchy in terrorist networks: Some case studies." *Lecture Notes in Computer Science 5075*: 477–489.

Microsoft Netscan. 2006. http//netscan.research.microsoft.com

Nanavati, A. A., Gurumurthy, S., Das, G., Chakraborty, D., Dasgupta, K., Mukherjea, S., and Joshi, A. 2006. "On the structural properties of massive telecom graphs: Findings and implications." In *Proceeding of the fifteenth ACM CIKM Conference on Information and Knowledge Management*.

Newman, M. 2001. "The Structure of Scientific Collaboration Networks." In *Proceedings of the National Academy of Sciences* 98: 4-9.

Newman, M. E. J. 2004. "Fast algorithm for detecting community structure in networks." *Physical Review E*, 69: 066133.

Newman, M. E. J. and Girvan, M. 2004. "Finding and evaluating community structure in networks." *Physical Review E*, 69: 026113.

Newman, M. E. J. 2004. "Detecting community structure in networks." *Eur. Phys. J.* B 38: 321-330.

Newman, N. E. J. 2005. "A Measurement of Betweenness Centrality Based on Random Walks." *Social Networks* 27: 39.

Newman, M. E. J. 2006. "Modularity and community structure in networks." In *Proceedings of the National Academy of Sciences* 103: 8577–8583.

Newman, M. E. J. 2006. Finding community structure in networks using the eigenvectors of matrices. *Physical Review E*, 74:036104.

Nilsson, N. 1998. *Artificial Intelligence*. A new synthesis, first edition, San Fransisco USA. Morgan Kaufmann Publishers.

Ogatha, H. 2001. "Computer Supported Social Networking for Augmenting Cooperation." *Computer Supported Cooperative Work*. Kluwer Academic Publisher 10: 189-209.

Onnela, J., P., Saramaki, J., Hyvonen, J., Szabo, G., Lazer, D., Kaski, K., Kertesz, J. and Barabasi. A., L. 2007. "Structure and tie strengths in mobile communication networks." In *Proceedings of the National Academy of Sciences of united State of America* 104 (18): 7332-7336.

Onnela, J. P., Saramaki, J., Hyvonen, J., Szabo, G., Menezes, M. A., Kaski, K., Barabasi, A. L. and Kertesz, J. 2007. "Analysis of a large-scale weighted network of one-to-one human communication." *New Journal of Physics* 9 (6) 179.

Palla, G., Barabasi, A. and Vicsek, T. 2007. "Quantifying social group evolution." *Nature* 446: 664-667.

Pentland, A. 2006. "Collective intelligence." *IEEE Computational Intelligence Magazine* 1: 9-12.

Pentland, A. 2007. "Automatic mapping and modeling of human networks" *Physica A: Statistical Mechanics and its Applications, Elsevier* 378: 59-67.

Popescul, A. and Ungar, L. H. 2003, "Structural Logistic Regression for Link Analysis." In *Proceedings of KDD Workshop on Multi-Relational Data Mining*, 2003.

Popescul, A. and Ungar, L. H. 2004. "Cluster-based Concept Invention for Statistical Relational Learning." In *Proceedings of ACM International Conference on Knowledge Discovery and Data Mining,* 22-25.

Qin, S., Qian, W. and Zhou, A. 2005. "Adaptively Detecting Aggregation Bursts in Data Streams." In *Proceedings of 10th International Conference on Database Systems for Advanced Applications*.

Raftery, A. E. and Akman, V. E. 1986. "Bayesian analysis of a Poisson process with a change-point." *Biometrika* 73: 85–89.

Rattigan, M. J. and Jensen, D. 2005, "The Case for Anomalous Link Detection." In *Proceedings of the 4th international workshop on multi-relational mining*, 69-74.

Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N. and Barabasi, A. L. 2002. "Hierarchical organization of modularity in metabolic networks." *Science* 297(5586):1551–1555.

Reality Mining. 2009. Massachusetts Institute of Technology. http://reality.media.mit.edu/

Ritov, Y., Raz, A. and Bergman, H. 2002. "Detection of onset of neuronal activity by allowing for heterogeneity in the change points." *Journal of Neuroscience Methods* 122: 25–42.

Sack, W. 2000. "Conversation Map: A text-based Usenet Newsgroup Browser." In *Proceeding of ACM Conference on Intelligent User Interfaces*, 233-240.

Salmenkivi M. and Mannila, H. 2005. "Using markov chain monte carlo and dynamic programming for event sequence data." *Knowledge and Information Systems* 7 (3) 267–288.

Sarkar, P. and Moore, A. 2005. "Dynamic Social Network Analysis using Latent Space Models." In *Proceeding of SIGKDD Explorations: Special Edition on Link Mining*.

Schnegg, M. 2006. "Reciprocity and the Emergence of Power Laws in Social Networks." *Int. J. Mod. Phys*. C 17 (8): 1-11.

Shi, X., Tseng, B. and Adamic, L. A. 2007. "Looking at the blogosphere topology through different lenses." In *Proceedings of the International Conference on Weblogs and Social Media*.

Stern, C. 1981. "Estimation of the Mean of Multivariate Normal Distribution." The *Annals of Statistics* 9: 1135-1151.

Taskar, B., Wong, M. F., Abbeel, P. and Koller, D. 2004. "Link prediction in relational data." In *Proceedings of Neural Information Processing Systems*, 13-18.

Taylor, J. S. and Cristianini, N. 2004. *Kernel Methods for Pattern Analysis*. Cambridge University Press, New York, NY, USA.

Teng, W. and Chou, M. 2007. "Mining communities of acquainted mobile users on call detail records." In *Proceedings of the 22nd Annual ACM Symposium on Applied Computing*.

Viegas, F. and M. Smith, M. 2004. Newsgroup Crowds and AuthorLines". In *Proceedings of the Hawaii International Conference on System Science*.

Vlachos, M., Meek, C., Vagena, Z. and Gunopulos, D. 2004. "Identifying similarities, periodicities and bursts for online search queries." In *Proceedings of the 2004 ACM SIGMOD International Conference on Management of Data*.

Vlachos, M., Wu, K., Chen, S. and Yu, P. 2005. "Fast Burst Correlation of Financial Data." In *Proceedings of 9th European Conference on Principles and Practice of Knowledge Discovery in Databases*, 368-379.

Wang, M., Madhyastha, T. M., Chan, N. H., Papadimitriou, S. and Faloutsos, C. 2002. "Data mining meets performance evaluation: Fast algorithms for modeling bursty traffic." In *Proceedings of the Eighteenth International Conference on Data Engineering*.

Wang, X. and McCallum, A. 2006. "Topics over Time: A Non-Markov Continuous-Time Model of Topical Trends." In *Proceeding of 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Wasserman. S and Faust. k., 1994. *Social Network Analysis: Methods and Applications*. Cambridge University Press.

Watts, D., and S. Strogatz, 1998. "Collective Dynamics of Small-World Networks. *Nature* 393: 440-442.

Welser, H. T., Gleave, E., Fisher, D., and Smith, M. 2007. "Visualizing the signatures of social roles in online discussion groups." *Journal of Social Structure* 8: 1-32.

Worsley, K. J. 1979. "On the likelihood ratio test for a shift in location of normal Populations." *Journal of the American Statistical Association* 74: 365-367.

Yamanishi, K. and Takeuchi, J. 2006. "A unifying framework for detecting outliers and change points from time series." *IEEE transactions on knowledge and data engineering* 18 (4) 482-492.

Yang, T. Y. and Kuo, L. 2001. "Bayesian binary segmentation procedure for a Poisson process with multiple change points." *Journal of Computational and Graphical Statistics* 10: 772–785.

Yuan, Z., Jia, Y. and Yang, S. 2007. "Online Burst Detection Over High Speed Short Text Streams." In *Proceedings of International Conference on Computational Science*, 717-725.

Zamora-López, G., Zlatić, V., Zhou, C., Štefančić, H. and Kurths, J. 2008. "Reciprocity of networks with degree correlations and arbitrary degree sequences." *Phys. Rev*. E **77**: 016106.

Zhang, H. and Dantu, R. 2008. "Quantifying the presence for phone users." In *Proceeding of the fifth IEEE Consumer Communications & Networking Conference*.

Zhao, Q. and Mitra, P. 2007. "Event Detection and Visualization for Social Text Streams." In *Proceedings of International Conference on Weblogs and Social Media*.

Zhou, D. and Scholkopf, B. 2004, "A regularization framework for learning from graph data" In *Proceedings of Workshop on Statistical Relational Learning at International Conference on Machine Learning.*

Zhu, Y. and Shasha, D. 2003. "Efficient elastic burst detection in data streams." In *Proceedings of the ninth ACM SIGKDD international conference***,** 336 – 345