# Socioscope: Human Relationship and Behavior Analysis in Social Networks

Huiqi Zhang, Ram Dantu, *Member, IEEE*, and João W. Cangussu, *Member, IEEE*

***Abstract*—In this paper, we propose a socioscope model for social-network and human-behavior analysis based on mobile-phone call-detail records. Because of the diversity and complexity of human social behavior, no one technique will detect every attribute that arises when humans engage in social behaviors. We use multiple probability and statistical methods for quantifying social groups, relationships, and communication patterns and for detecting human-behavior changes. We propose a new index to measure the level of reciprocity between users and their communication partners. This reciprocity index has application in homeland security, detection of unwanted calls (e.g., spam), telecommunication presence, and product marketing. For the validation of our results, we used real-life call logs of 81 users which contain approximately 500 000 h of data on users' location, communication, and device-usage behavior collected over eight months at the Massachusetts Institute of Technology (MIT) by the Reality Mining Project group. Also, call logs of 20 users collected over six months by the University of North Texas (UNT) Network Security team are used. The MIT and UNT data sets contain approximately 5000 callers. The experimental results show that our model is effective.**

***Index Terms*—Change points, reciprocity index, social groups, social networks, social relationships, socioscope.**

## I. Introduction

A SOCIAL network is defined as a set of actors (individuals) and the ties (relationships) among them [1]. These relational ties and actors compose the fundamental interests in social networks. Also, a social group can be defined as a set of people who have common interests, i.e., like the same subjects, share their experience, express similar ways of thinking and reacting, share the same opinions, do similar things, and have the same goals. They actively exchange information. In the presence of new events, they discuss with each other to decide what to do.

Manuscript received October 2, 2009; revised June 22, 2010; accepted September 26, 2010. Date of publication March 17, 2011; date of current version October 19, 2011. This work was supported by the National Science Foundation under Grants CNS-0627754, CNS-0516807, CNS-0619871, and CNS-0551694. Any opinions, findings, conclusions, or recommendations expressed in this paper are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. This paper was recommended by Associate Editor L. Rothrock.

H. Zhang is with the Department of Computer Science and Engineering, University of North Texas, Denton, TX 76203 USA (e-mail: huiqizhang@my.unt.edu).

R. Dantu is with the Department of Computer Science and Engineering, University of North Texas, Denton, TX 76203 USA and also with the Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: rdantu@unt.edu).

J. W. Cangussu is with the Microsoft Corporation, Redmond, WA 98052-6399 USA (e-mail: cangussu@microsoft.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSMCA.2011.2113335

The modern telecommunication and Internet technologies, such as mobile communications and social media, unite people around the world into a wide-area social network (WASN). We may communicate with our family members, relatives, friends, and colleagues and find new friends who have the same interests as we do at any time and almost anywhere. In a WASN, people form groups or clusters based on interests, goals, etc. Since the mobile phones have become an important tool of modern human daily life, telecommunication patterns may reflect different human relationships and behaviors, and changes in the telecommunication pattern may expose signs of social relationships and behavior changes. For example, the calling patterns of a person with his/her friends differ from those with spammers.

Organizations or individual elements may be interested in different social-network properties. For example, people in homeland security-related departments are interested in particular groups of persons such as terrorists, robbers, and other groups that present security risk. Businesspersons want to know which groups of people are interested in their products. Network designers and operators want to know overall users' distributions and patterns to efficiently and effectively use and distribute resources and enhance the quality of services. Unfortunately, almost all existing social-network research has focused on overall social-network structures and properties such as clusters and centrality. These research efforts lack analysis for one-to-one or one-to-many relationships and behaviors in detail—necessary when interested in special groups or clusters of people. These detailed features of human relationships are more important for detecting terrorists, spam, and user preferences. Because of human social behavior's diversities and complexities, applying one technique will not detect the many different features of human social behaviors. Therefore, we use multiple probability and statistical methods, integrating them for social-network and human-behavior analysis from macrolevels to microlevels. We propose a socioscope model for analyzing the properties of social structures and human behavior; for quantifying and measuring interpersonal relations in groups; for detecting change points, unusual consumption events, and opt-in bursts; and for identifying willingness levels of users 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, for discovering social groups and events, for quantifying relationships, and so on. The socioscope is extensible; new tools can be added as needed. By zoom-in, we may use multiple scales to analyze social-group member

behavior up to the one-to-one relationship. By zoom-out, we may analyze general social-network structures and properties.

The remainder of this paper is organized as follows. In Section II, we briefly review the related work. In Section III, we describe the model, architecture, and components for the socioscope. We describe the methods for quantifying social groups and reciprocity index, event detection, and pattern recognition in Sections IV–VI, respectively. We perform the experiments and validation with actual call logs and discuss the results in Section VII. Finally, we have the conclusions in Section VIII.

## II. Related Work

The work proposed in Sections III–VI refers to the social-network structure and the relationships among the people on the network, along with event detection based on their call records. Existing work on these topics is described in the next sections.

### A. Social Groups

The study of social networks has been applied in modern sociological studies for some time. The major applications focus on tasks, such as measuring interpersonal relations in groups and describing the properties of social structures and individual social environments [1].

Ogatha [2] discusses the strength of social relations between two persons which they measured with e-mail conversation. The relation is strong if the e-mail between two persons is exchanged frequently, recently, and reciprocally, and Ogatha uses a formula for the strength, which is a function of user-determined importance weights and the number of received and sent e-mails. Newman [3] investigates the automatic detection methods of the subgroup structure of social networks based on expected density and edge betweenness. Broder *et al.* [4] use a block technique, which focuses on the pattern of connectivity to find clusters of relationships that might be hidden. Doreian *et al.* [5] propose the generalized block modeling method to enable partitions of smaller more precise block types and predefined block structures based on attribute data. In [6] and [7], the studies focus on identifying tightly connected clusters or communities within a given graph and inferring potential communities within a network based on the density of linkage. Newman uses the self-identified nature of the online communities to study the relationship between different newsgroups on Usenet [8]. In [9]–[11], social-network evolution is studied as its members' attributes change. Backstrom *et al.* [12] perform a large-scale empirical analysis of social-network evolution in which interactions between people are inferred from time-stamped e-mail headers. The social-network-evolution model of topics over time is proposed in [13].

The social-network analysis and social clusters of the aforementioned work are mainly based on blogs, e-mails, or the World Wide Web [14]–[16]. In [17] and [18], the structure and tie strength of mobile-phone-call graphs are investigated. Kurucz *et al.* [19] apply the spectral clustering method to telephone-call graph partition. Teng and Chou [20] identify the communities of mobile users on call-detail records using a triangle approach. Palla *et al.* [21] have developed an algorithm based on clique percolation to investigate the time dependence of 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.* [22] present a 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). The MRQAP is a standard technique to analyze social-network data, to discover behavioral characteristics of friendship using factor analysis, and to predict satisfaction based on behavioral data. Hidalgo and Rodriguez-Sickert [23] study the stability of social ties by defining and measuring the ties' persistence. They show that the persistence of ties and perseverance of nodes depend on the degree, clustering, reciprocity, and topological overlap. Dasgupta *et al.* [24] propose a spreading activation-based technique to predict potential churners by examining a current set of churners and its underlying social network. Candia *et al.* [25] investigate the spatiotemporal anomalies of calls and patterns of calling activity using standard percolation theory tools. They report that the interevent time of consecutive calls is heavy tailed. Almost all of the aforementioned research focuses on the general features of social networks and social groups.

Pentland [26], [27] uses the mobile phones programmed, electronic badges, and microphones as a socioscope to sense and capture human behavioral data (location, proximity, and body motion). These behavioral data are then used to analyze the characterization of group distribution and variability and the conditional probability relationships between individual behaviors and to focus on human relationship analysis based on physical distance proximity [22]. Eagle [28] extends the approach in [26] and [27] to study a variety of human cultures as a culture lens.

Eagle and Pentland [29] identify the structure inherent in daily behavior with models to analyze, predict, and cluster multimodal data from individuals and communities within the social network of a population which are focused on temporal location data. They represent this behavioral structure by the principal components of the complete behavioral data set, a set of characteristic vectors which are termed as eigen behaviors (e.g., {Home, Elsewhere, Work, No Signal, Off}). In this model, an individual's behavior over a specific day can be approximated by a weighted sum of his or her primary eigen behaviors. Eagle *et al.* [30] propose a model by combining the average physical distance to travel, clustering coefficient, ego density, and call frequencies to provide comparative insight into the human and social behavior between urban and rural communities. They found that the individuals living in urban areas tend to communicate almost 50% more than the individuals living in rural areas.

Eagle [31] discusses the need for a set of standardized protocols for behavioral data acquisition and usage. Eagle [32] presents a system that enables people to earn small amounts of money by completing simple tasks which include translation, transcription, and surveys on their mobile phone for corporations who pay them.

Gallagher *et al.* [33] propose an approach for the classification in partially labeled networks focused on predicting node labels in a large graph. They introduced "ghost edges" by judiciously adding edges between nodes and proposed to bypass

all the activation-spreading methods to solve this problem. Koutsourelakis [34] proposes a nonparametric Bayesian framework in which the size of the model, i.e., the number of clusters, can adapt to the available data to discover unsupervised groups in relational data sets. Carreras *et al.* [35] propose a model based on an extension of the eigenvector centrality principle for analyzing the spread of epidemics in a disconnected mobile network. They define a new connectivity matrix used to evaluate the eigenvector centrality of the various nodes.

Wang *et al.* [36] propose the WR-KMeans model for the instant message clustering. WR-KMeans is a variant of the standard $k$-means algorithm. The WR-KMeans model extends the traditional term frequency–inverse document frequency (TF-IDF) model of conversations by relevant words. WR-KMeans measures the similarity between conversations according to a cosine measure. Chaintreau *et al.* [37] analyzed the diameter of the mobile network. They found that the network diameter generally varies between three and six hops for networks containing 40 up to a hundred nodes for sparse and dense networks. Hui *et al.* [38] propose three distributed community detection algorithms, the SIMPLE, k-CLIQUE, and MODULARITY. The three methods are based on centrality and clustering coefficient measurements.

Eagle and Pentland [39] present ethnographic studies of device usage, relationship inference, individual behavior modeling, and group behavior analysis. They used a Gaussian mixture model to detect patterns in proximity between users and correlate them with the type of relationship. Lawrence *et al.* [40] discover copresence communities between a group of individuals by mining a set of raw copresence events captured by Bluetooth-enabled mobile phones. They used the well-known concept formation algorithms for clustering. Eagle [41] presents ethnographic studies of devices usage, self-report data validation, relationship inference, individual behavior modeling, and group behavior analysis. He discovered regular and predictable rules and structure in the behavior of individuals, teams, and organizations by machine learning methods using cell tower and bluetooth information.

Kobashikawa *et al.* [42] propose a fuzzy algorithm for group decision making, considering the finite discriminating abilities of real decision makers. In [43], the partitioning dynamic clustering methods for interval-valued data were proposed based on adaptive quadratic distances. Yager [44] introduces the idea of fuzzy relationships in modeling weighted social relational networks. Maulik *et al.* [45] propose a modified differential evolution-based fuzzy c-medoids clustering algorithm of categorical data. The algorithm combines both local and global information using adaptive weighting.

In [22], [26]–[32], and [39]–[41], the approaches for social-group discovery are mainly based on proximity (physical distance), temporal, and location information. In [2]–[21], [23], and [33]–[38], the approaches for social-group discovery are mainly based on *centrality* and *clustering coefficient* methods in graph theory. All of the aforementioned previous work mainly focuses on general social-network structures.

In this paper, we focused on quantifying individual social groups up to the one-to-one relationship using a probability model, *affinity* [46] which is virtual distance, based on mobile-phone call-detail records. We used *affinity* to measure the similarity between probability distributions and for quantifying the social ties' strength between actors in groups. This paper differs from previous work on measurements which focused on general structures of social networks. Because phone calls are stochastic processes, we argue that it is more suitable to use probability affinity to quantify social relationships.

### B. Dyads and Reciprocity Index

Gouldner [47] proposes an index of mutuality to measure the tendency toward mutuality by the probability of mutual choices between two actors. Katz *et al.* [48], [49] propose an index to measure the tendency for mutuality, which compares 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 they calculate a z-score. Schnegg [50] finds that, if he includes the reciprocity's effect and the scaling exponent, which are negatively correlated in simulations of a growing network, the degree distributions are much closer to those empirically observed. Garlaschelli and Loffredo [51] propose a framework in which the occurrences of mutual links depend on conditional connection probabilities according to their actual degree of correlation between mutual links. Zamora-López *et al.* [52] report that one- and two-node degree correlations are 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.* [53] investigates the lattice reciprocity mechanisms and interprets the onset of lattice reciprocity as a thermodynamic phase transition to enhance the evolutionary survival of the cooperative phenomena in social networks. Hogan and Fisher [54] find that reciprocity in e-mail behavior differs between multirecipient and dyadic mail.

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

### C. Change-Point Detection

A large amount of work on change-point detection exists. Baron [55] proposes efficient online and offline nonparametric algorithms for detecting the change point, based on histogram density estimators. Baron and Granott [56] develop schemes for detecting early change points and frequent change points. The schemes possess a number of desired properties, including distribution consistency, which implies the convergence of small-sample change-point estimators. The aforementioned methods are applied to the temperatures, climate, and

software-engineering quality-control data [57]. Raftery and Akman [58] develop a Bayesian approach for detecting a single change point at an unknown time of a Poisson process. Yang and Kuo [59] propose a Bayesian binary segmentation procedure for detecting multiple change points for a Poisson process. Ritov *et al.* [60] apply the Bayesian change-point method to neural data under the assumption of inhomogeneous Poisson process. Carlin *et al.* [61] proposes a changing linear-regression model and obtains the desired posterior densities by an iterative Monte Carlo method. Lund and Reeves [62] propose the revision of a two-phase linear regression model to apply to climate data. Hawkins [63] develops a procedure for detecting change points of a series of varying normal means under the assumption of the known variance. Worsley [64] extends the method in [63] for a series of varying normal means with unknown variance. These procedures are performed based on a likelihood ratio test. Chen and Gupta [65] use a binary procedure combined with the 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.* [66] applies a reversible-jump Markov chain Monte Carlo simulation to estimate the variance change points of activation patterns from electromyographic data with an assumption of the data being a zero mean. The variance is modeled by a step function. Chu and Zhao [67] apply a Bayesian approach to detect the change points in a time series of annual tropical cyclone counts under the assumption of a Poisson process with gamma distribution. Fearnhead [68] performs direct simulation from the posterior distribution of multiple change-point models with an unknown number of change points based on recursions. The class of models assumes that parameters associated with segments of data between successive change points are independent. Yamanishi and Takeuchi [69] propose a scheme for detecting outliers and change points for nonstationary time series. The main feature of this scheme is that an outlier is first detected by the model learned in the first stage, which repeats the learning process twice, and the change point is detected by the learned model in the second one. Yamanishi and Takeuchi apply their scheme to autoregressive models and an sequential discounting auto regression (SDAR) learning algorithm as learning modules. This approach presents a variant of a maximum-likelihood method for online discounting learning of that model, which is adaptive to nonstationary time series. Erdman and Emerson [70] apply a Bayesian method to analyze change points for the segmentation of microarray data with the implementation of the Bayesian change-point method in linear time.

### *D. Unusual Consumption Event Detection*

Almost all previous work on event detection is based on text, data stream, and video. Brants and Chen [71] propose a method based on an incremental TF-IDF model. Extensions include the generation of source-specific models, similarity score normalization based on document-specific averages, source-pair specific averages, term reweighting based on inverse event frequencies, and segmentation of the documents. Chen *et al.* [72] examine the effect of a number of techniques, such as part of speech tagging, similarity measures, and an expanded stop list on performance. Kumaran and Allan [73] use text classification techniques and named entities to improve performance. Li and Croft [74] propose a novelty detection approach based on the identification of sentence level patterns. Li *et al.* [75] propose a probabilistic model to incorporate both content and time information in a unified framework. This approach gives new representations of both news articles and news events. They also explore in 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 [76] propose detecting events by combining text-based clustering, temporal segmentation, and graph cuts of social networks. In this method, each node represents a social actor, and each edge represents a piece of text communication that connects two actors. In [77] and [78], the authors propose a conceptual model-based approach by the use of domain knowledge and named-entity-type assignments. They show that a classical cosine similarity method fails for the anticipatory event detection task. Luo *et al.* [79] proposes an online new event detection (ONED) framework, which includes the following features: a combination of indexing and compression methods to improve the document processing rate and a resource-adaptive computation method to maximize the benefit that can be gained from limited resources. The *ONED* framework also identifies new events to be 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 and when implicit citation relationships need to be created among all the documents and used to compute the importance of document sources.

Guralnik and Srivastava [80] propose an iterative algorithm and use a likelihood criterion to segment a time series into piecewise homogeneous regions to detect those change points, which are equivalent to the events defined by the change points, and to evaluate the change points within highway traffic data. Kleinberg [81] uses an infinite automaton, in which bursts are state transitions, to detect burst events in text streams. He conducts his experiments with e-mails and research papers. Keogh *et al.* [82] use a suffix tree to encode the frequency of observed patterns and apply a Markov model to detect patterns in the symbol sequence. Salmenkivi and Mannila [83] use piecewise constant intensity functions to represent continuous intensity functions using a combination of Poisson models and Bayesian estimation methods and use dynamic programming methods to find events. Ihler *et al.* [84] use a time-varying Poisson-process model and statistical estimation techniques for unsupervised learning in the context. They apply this model to freeway traffic data and building access data. Yang *et al.* [85] propose a model to identify the event evolution relationships between events in an incident by using the event timestamp, event content similarity, temporal proximity, and document distributional proximity.

We propose and investigate an inhomogeneous Poisson and inhomogeneous exponential distribution model to detect events, and we illustrate how to learn such a model from data to both characterize normal behavior and detect anomalous events based on call-detail records. The main difference between

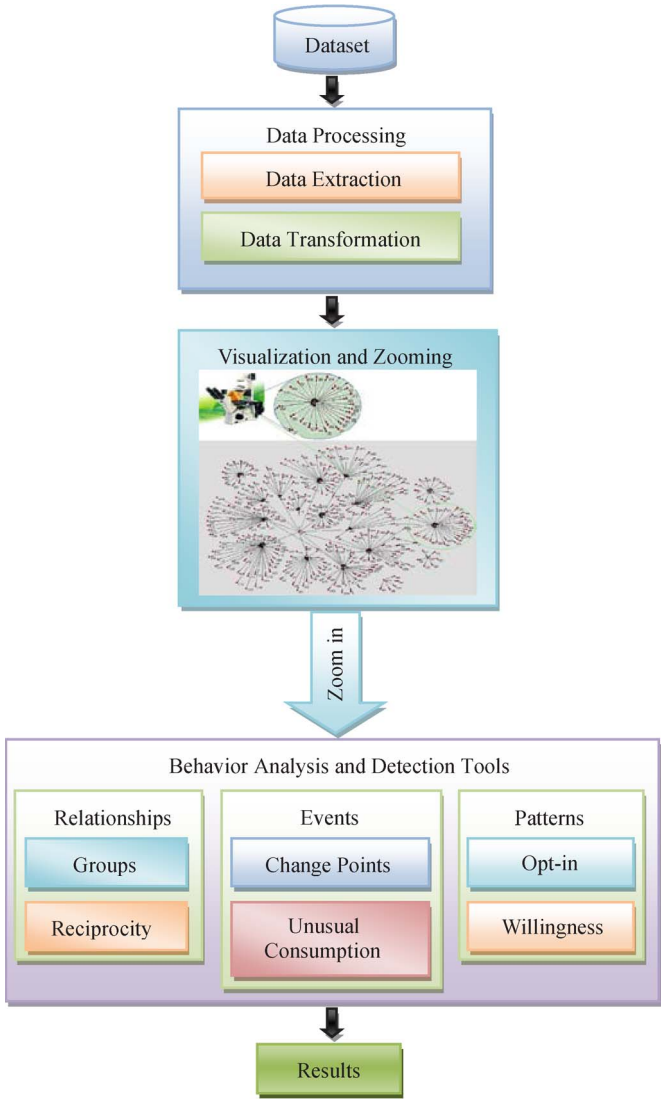


Fig. 1. Architecture of the socioscope.

call-detail records and text or Web site data is that call-detail records have no content. Therefore, it is more difficult to detect events hidden in these records. We can only use external information such as the time of call initiation, number of calls within a period of time, call duration, incoming calls, outgoing calls, and caller location. We use maximum-likelihood estimation to estimate the rates and thresholds of the number of calls and call duration [86].

## III. Overview of the Proposed Approach

Socioscope provides a platform for analyzing the properties of social-network structures and human social behavior, for measuring interpersonal relations in groups, and for discovering special events based on human telecommunication patterns. Next, we describe the socioscope model and its components.

As presented in Fig. 1, the socioscope model consists of a number of components, including data extraction and transformation; network visualization; and zoom, scale, and several analysis tools used for analyzing network structures, discovering and quantifying social groups and events, quantifying relationships, etc. The socioscope model is extensible. New tools can be added as we identify additional features and uses. The model is composed of the three layers briefly described hereinafter.

1) *Data Processing*: This layer consists of two components—*Data Extraction and Data Transformation*. In data extraction, related information is extracted from raw data sets and then transformed in data transformation into the required data format for visualization and analysis.
2) *Visualization and Zooming*: An open-source visualization tool is used for drawing the social networks. Using zooming-in levels, we may use multiple scales to analyze social-group member behavior up to the one-to-one relationship. Using zooming-out levels, we may analyze general social-network structures and properties.
3) *Behavior Analysis and Detection Tools*: This layer, the core of the model, consists of four components—*Quantifying Social Groups* and *Reciprocity, Change Point Detection* and *Unusual Consumption Detection, Opt-in Pattern Detection*, and *Willingness Level inference*.

We describe these components and solutions in detail next.

## IV. Quantifying Social Groups

Most social-network research and social-relationship analysis are based on blogs, e-mails, or data drawn from World Wide Web [14]–[16]. Because mobile phones have become the main communication media for many people in recent years, some researchers' interests in social networks concentrate on social-relationship analysis based on call-detail records [17]–[32]. Almost all this research focuses on general structures and properties for social networks. In real life, we are usually interested in the behaviors and quantitative relationships between some particular groups of people. For example, in marketing, if someone buys a product, his family members and friends are likely to have the same interest in buying the same or a similar item. They may also have a similar income level although we do not know how much they earn. Therefore, a business may find potential buyers by studying social groups. Another important application for social groups is national security. For example, if a person is a terrorist or a robber, his intimate friends or socially close communication partners are likely (though not necessarily) to be terrorists or robbers too, since most law-abiding persons do not want to have friends who are terrorists or robbers. Another application might be used to quantify a group's telecommunication presence. On different days and at different times, people usually will communicate with different groups of people. For example, most prefer to communicate with colleagues during work hours and to communicate with family members, relatives, and friends during nonwork time. Furthermore, during busy hours, most would like to have only necessary communications with our socially close members such as family members, bosses, and friends. Additionally, we may enhance detecting unwanted calls (e.g., spam) by social-group analysis. For example, spammers are definitely socially far from us. If we are not sure that an incoming call from a socially far member is spam or not, the system may not let

the phone ring and will forward the calls to a voice mail box automatically.

The approach proposed here for the social-group identification relies first on a computation of a reciprocity index. We then use the index to compute the affinity between two users. Finally, we use the affinity to define the socially related groups. These steps are presented in the next sections.

### A. 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 similar acts [1], [48]. In this paper, our interest is to investigate how people use technology to construct social relationships. We focus on the measure of mediated interactions considering the media used to interact. To investigate how people interactively construct their social relationships, we focus 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. The buyers receive better products and service, and the sellers earn more money. Therefore, the reciprocity relation is one of the keys to business success. Similarly, we may enhance detecting unwanted calls (e.g., spam) by reciprocity analysis. For example, spammers definitely do not receive responses from us. Thus, there is little or no reciprocity in this relationship.

Most social relationship research focuses on the collection of dyads in social networks [1]. We propose a new reciprocity index that differs from the previous work as it also accounts for the communications' time and duration.

The dyadic relationship in a social network is the collection of dyads (an unordered pair of nodes (actors) and arcs (ties) between the two nodes). There are $(n \times (n-1))/2$ 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 [30], the authors propose the index of mutuality $\rho_{kp}$. This index focuses on the probability of a mutual choice between two actors $i$ and $j$

$$P(i \to j \& j \to i) = P(i \to j) \times P(j \to i | i \to j). \quad (1)$$

$P(j \to i | i \to j)$ can be considered as consisting of two parts: $P(j \to i)$ and a fraction, denoted by $\rho_{kp}$ of the probability $P(j \xrightarrow{not} i)$ [48]. $\rho_{kp}$ is zero if there is no tendency toward mutuality and one if there is a maximal tendency toward mutuality. A negative value of the index indicates a tendency away from mutuality, toward asymmetry and nulls (referred to as antireciprocation). There are two kinds of $\rho_{kp}$, *fixed choice* and *free choice*. The fixed choice assumes that all actors make the same number of choices, and the estimate of $\rho_{kp}^{\text{fixed}}$ is computed by

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

where $n$ is the number of nodes, $M$ is the observed number of mutualities, and $c$ is the number of choices [48].

The free choice allows different numbers of choices, and the estimate of $\rho_{kp}^{\text{free}}$ is computed by

$$\hat{\rho}_{kp}^{\text{free}} = \frac{2(n-1)^2 M - S^2 + S_2}{S(n-1)^2 - S^2 + S_2} \quad (3)$$

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 [48].

The indices in [48] focus on the probability of mutual choice of nodes in a graph. The authors neither deal with the frequencies of communication transaction nor consider the reciprocity time between two nodes. The reciprocities of once and multiple times are the same thing, i.e., there is a link between two nodes no matter how many times people contact each other. However, in the real world, the relationship between two persons is different if their reciprocity frequencies and response time are different.

In mobile-phone social networks, actor $i$ and actor $j$ may call each other multiple times. Reciprocity reflects their relationship in a period of time. The aforementioned 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 communication frequency. We propose a reciprocity index $\rho_{a \leftrightarrow b}$ which does measure the tendency of reciprocity for actors $a$ and $b$ in a group.

Fig. 2 shows the reciprocity relation between phone user29 and his communication partner 349 where $m$, $h$, $d$, and the numbers inside the boxes above the arrows indicate minute, hour, day, and call duration, respectively.

Suppose that the number of phone call arrivals is a Poisson process; this has been shown by Bregni *et al.* [87]. Then, the probability of no arrivals in the interval $[0, t]$ is given by

$$P(\tau > t) = e^{-\lambda t} \quad (4)$$

where $\lambda$ is the arrival rate and $\tau$ is interarrival time. The occurrence of at least one arrival between 0 and $t$ is given by

$$P(\tau \le t) = 1 - e^{-\lambda t}. \quad (5)$$

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

$$\begin{aligned} &P(a \to b \,\&\, b \to a) \\ &\quad = P(a \to b) P(b \to a | a \to b) \\ &\quad = P(a \to b) \left[ P(b \to a) + \rho_{a \leftrightarrow b} P(b \xrightarrow{not} a) \right] \\ &\quad = (1 - e^{-\lambda_a t_i}) \left[ \left( 1 - e^{-\lambda_b (t_j - t_i)} \right) + \rho_{a \leftrightarrow b} e^{-\lambda_b (t_j - t_i)} \right]. \end{aligned} \quad (6)$$

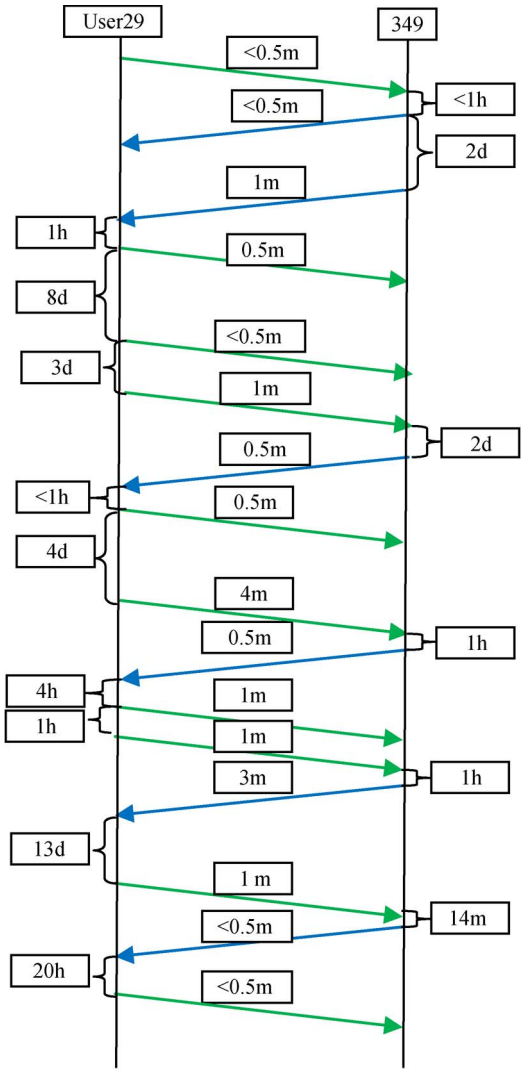


Fig. 2. Event flow chart for phone user29 with his partner 349.

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

$$E(R|\rho_{a\leftrightarrow b}) = S(1-e^{-\lambda_a t_i}) \times \left[\left(1-e^{-\lambda_b(t_j-t_i)}\right) + \rho_{a\leftrightarrow b}e^{-\lambda_b(t_j-t_i)}\right]. \quad (7)$$

After rearranging the terms, we have

$$\rho_{a\leftrightarrow b} = \left[R - S(1-e^{-\lambda_a t_i})\left(1-e^{-\lambda_b(t_j-t_i)}\right)\right] / S(1-e^{-\lambda_a t_i})e^{-\lambda_b(t_j-t_i)} \quad (8)$$

where $R$ is the observed number of reciprocity.

$\rho_{a\leftrightarrow b}$ is zero if there is no tendency toward reciprocity and one if there is a maximal tendency toward reciprocity.

We compute the reciprocity indices by (8) for the call-log data. In this paper, we define the reciprocity time interval $t_j - t_i$ as 24 h, i.e., the returned calls or messages within a 24-h period, to compute the reciprocity index. (This is only an example to choose $t_j - t_i$. We may adjust this parameter to any reasonable length of time.)

### B. Social-Group Identification

Groups correspond to data clusters. Cluster analysis concerns a set of multivariate methods for grouping data variables into clusters of similar elements.

One limitation in [22], [26]–[32], and [39]–[41] is that the authors do not deal with the relationships of persons whose physical distances are very far away. We may keep relationships with our family members, friends, coworkers, etc. not only based on close proximity (face to face) but also based on telecommunication technologies even when they are several thousands of miles away.

The widely used approaches in [2]–[21], [23], and [33]–[38] for social-group discovery are mainly based on *centrality* and *clustering coefficient* methods in graph theory. These approaches do not distinguish the kind of relationships of persons in a group. In the real world, the relationship between each pair in a group is different.

We use a microscopic analysis strategy to further quantify social relationships of persons in a group and categorize them as socially close, near, and far members. Our approach can further analyze the human relationships in the same group in more detail. *Another main difference of the methods between our approach and existing approaches is how we deal with reciprocity between two people*.

Our approach is mainly based on transactional (calling patterns, number of calls, call duration, reciprocity, and affinity), temporal, and location information of people. Their physical distances do not matter since we may build relationships with people worldwide by telecommunication technologies.

In this paper, we use probabilistic models for the classification of variables by their *affinity* [46].

We can represent the observed data in a 2-D matrix, where the rows describe the data units and the columns describe the categorical variables. Usually, empirical clustering models are used to analyze such data. In these instances, in the first step, one chooses 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 defines an aggregation criterion for merging similar clusters of elements, and in the third step, one assesses the validity of the clustering results. In this paper, we apply a more appropriate probabilistic model. In the first step, we apply a probabilistic similarity coefficient, i.e., *affinity* (measured in a probability scale), instead of simple or basic similarity coefficients. In the second step, we define an aggregation criterion for merging similar clusters of elements. In the third step, we use internal validation to assess the validity of the clustering results, i.e., similarity coefficients comparing a classification with the original data sets.

Affinity measures the similarity between probability measures. Because our problem belongs to discrete events, we only consider finite event spaces. Let

$$S_N = \left\{P = (p_1, p_2, \ldots p_N) | p_i \geq 0, \sum_{i=1}^{N} p_i = 1\right\} \quad (9)$$

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

$$d_H^2(P,Q) = \frac{1}{2}\sum_{i=1}^{N}(\sqrt{p_i} - \sqrt{q_i})^2 \tag{10}$$

$d_H^2(P,Q) \in [0,1]$, $d_H^2(P,Q) = 0$ if $P = Q$ and $d_H^2(P,Q) = 1$ if $P$ and $Q$ are disjoint [26].

The affinity between probability measures $P$ and $Q$ is defined as

$$A(P,Q) = 1 - d_H^2(P,Q) = \sum_{i=1}^{N}\sqrt{p_i q_i} \tag{11}$$

$A(P,Q) \in [0,1]$, $A(P,Q) = 1$ if $P = Q$ and $A(P,Q) = 0$ if $P$ and $Q$ are disjoint [26].

*Proof:*

$$\begin{aligned} A(P,Q) &= 1 - d_H^2(P,Q) = 1 - \frac{1}{2}\sum_{i=1}^{N}(\sqrt{p_i} - \sqrt{q_i})^2 \\ &= 1 - \frac{1}{2}\sum_{i=1}^{N}(p_i - 2\sqrt{p_i}\sqrt{q_i} + q_i) \\ &= 1 - \frac{1}{2}\left(\sum_{i=1}^{N} p_i - 2\sum_{i=1}^{N}\sqrt{p_i q_i} + \sum_{i=1}^{N} q_i\right) \\ &= \sum_{i=1}^{N}\sqrt{p_i q_i}\ . \end{aligned} \tag{12}$$

For finite and discrete data, let $M(X,Y)$ be an $L \times N$ matrix, where $X$ represents the set of data units and $Y$ is a set of $N$ categorical variables. In this paper, $Y_j$ $(j = 1, \ldots, N)$ is a vector of frequencies. Thus, $Y_j$ may be represented by the $L$ coordinates $n_{ij}$ $(i = 1, 2, \ldots, L)$ denoting the frequency. We will refer to the $j$th column profile as the corresponding conditional vector with $n_{ij}/\sum_{i=1}^{L} n_{ij}$. 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 represents a partition of some random sample of subjects in $L$ classes. In this paper, $p_i = n_{ij}/\sum_{i=1}^{L} n_{ij}$. The column profiles have a major role as the similarity between variable pairs will be measured using an appropriate function, the affinity, of their profiles.

In our lives, we have relationships with small groups of individuals within our social network such as family members, relatives, friends, neighbors, and colleagues. Based on these social relationships, we divide the time of our day into working time (8 A.M.–5 P.M.) and nonworking time (5:01 P.M.–7:59 A.M.). Note that, because the data in our data set were collected from university students, professors, and staff members, work times may differ from the regular work time (8 A.M.–5 P.M.) which we reported in [88]. However, in this paper, we still use a regular work time for generalization. Furthermore, in the two time periods that we study, we divide our social network members into three categories: socially close members, socially near members, and socially far members.

1) *Socially close members*: The people with whom we maintain our strongest social relationship. Quantifying by phone calls, we receive more calls from socially close members. We tend to talk to them for longer periods of time. Immediate family members, intimate friends, and colleagues on the same team belong to this category.
2) *Socially near members*: These relationships are not as strong as those of family members, intimate friends, and colleagues on the same team. Sometimes, not always, we connect with each other and talk for considerably longer periods than we typically talk. We mostly observe intermittent frequency of calls from these people. Distant relatives, general friends, colleagues on a different team, and neighbors generally fall into this category.
3) *Socially far members*: These people have weaker relationships with us and with each other in social life. They call each other with less frequency. We seldom receive calls from them, and our talk with each other occurs for shorter periods of time than are typical of the other two groups.

In this paper, we use the 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. $Q = (q_{in}, q_{out}, q_{reci})$ 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 (1).

We compute the affinity between $P$ and $Q$ as follows:

$$A(P,Q) = \sum_i \sqrt{p_i q_i} \qquad \text{where } i \in \{in, out, reci\}. \tag{13}$$

We use the actual call-log data to compute the affinity values by (13). We define:

1) socially close members if $0.9 < A(P,Q) \leq 1$;
2) socially near members if $0.3 < A(P,Q) \leq 0.9$;
3) socially far members if $0 \leq A(P,Q) \leq 0.3$.

The validation results of the approach proposed earlier are presented in Section VII.

## V. Event Detection

Another important element that we can extract from call records is the occurrence of events. This capability could be used for detecting network attacks. To identify events in the call records, we first use a wavelet-denoising method to process the data, and then, we apply the modified method described in [57] for detecting change points based on the number of calls and call durations. These steps are described next.

Social-network structures and relationships dynamically change over time. Still, 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 to detect whether any changes have occurred. Change-point detection 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 has been an area of increasing interest to those studying a number of application types. Now, the automatic detection of change points by studying patterns of human behavior has recently attracted more attention. For instance, one important application of change-point detection is in the area of homeland security. For example, terrorists and robbers often attack in groups. The leader may communicate with group members by wireless phones to plan, coordinate, and command an attack. During this period of time, the group's calling patterns will differ from its usual patterns. Typically, there will be more calls. The group members, particularly the leader, may engage in longer talk time than is the usual case. Also, the group members may meet at some particular place, possibly even the attack target, for planning and attacking activities. When combined with other evidence, the change-point detection of calling patterns can prove useful to prevent security threats.

We use *change point* to refer to a large-scale activity that differs relative to normal patterns of behavior. To understand such data, we must both understand the patterns of the typical behavior and detect and extract information from the deviations from the typical behavior.

To achieve this, we combine the wavelet denoising and sequential detection methods to detect change points based on call-detail records. There is no content in call-detail records. It is important to remember that call-detail records contain no content. We can only use available information such as the time that a call is initiated, number of calls in a specific period, the duration of a call, whether the calls are incoming or outgoing calls, and the location of the caller.

### A. Wavelet Denoising

Generally, for the denoising, the wavelet scaling function should have properties similar to that of the original signal. The general wavelet-denoising procedure follows three steps: wavelet selection, threshold selection, and inverse wavelet transform (IWT), which are discussed in the following sections.

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

$$\gamma = \frac{\sum (X - \overline{X})(Y - \overline{Y})}{\sqrt{\sum (X - \overline{X})^2 (Y - \overline{Y})^2}} \tag{14}$$

where $\bar{X}$ and $\bar{Y}$ are the mean value of set $X$ and $Y$, respectively.

*2) Threshold Selection:* After a suitable wavelet basis function is chosen, the discrete wavelet transform (DWT) decomposition of a signal will compress the signal's energy 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 the wavelet coefficients representing the signal and removes the low-magnitude coefficients which represent noise.

The waveShrink method [89] is widely used to estimate signal $x$. The two commonly used shrinkage functions are *hard* and *soft thresholding* functions defined as

$$S_\delta^{\text{hard}}(x) = \begin{cases} x & |x| > \delta \\ 0 & |x| \le \delta \end{cases} \tag{15}$$

$$S_\delta^{\text{soft}}(x) = \begin{cases} x - \delta & x > \delta \\ \delta - x & x < -\delta \\ 0 & |x| < \delta \end{cases} \tag{16}$$

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

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

1) *Universal Threshold*: The universal threshold as defined in [90]

$$\delta_U = \sigma \sqrt{2 \log(n)} \tag{17}$$

where $\sigma$ is the standard deviation of the noise and $n$ is the sample length. The universal threshold uses a single threshold for all wavelet coefficients. One drawback is signals of short duration. Other methods may provide more accurate results.

2) *SURE Threshold*: The SURE threshold is based on Stern's unbiased risk estimator [91]. This method minimizes the SURE function to determine an optimal threshold. The SURE threshold is defined as

$$\delta_{SURE} = \min_\delta SURE\left(\delta, \frac{x}{\sigma}\right) \tag{18}$$

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 $\delta$ is the candidate threshold, $x_i$ is the wavelet coefficient, $n$ is the data size, and $m$ is the number of the data points less than $\delta$. This method has the sparse wavelet coefficient problem and is often combined with the universal threshold in a hybrid method.

3) *Hybrid Threshold*: The hybrid threshold method combines the universal and SURE threshold methods [90], [92]. This method uses the universal threshold if the signal-to-noise ratio is low with a sparse wavelet coefficient; otherwise, the method uses the SURE threshold method.

4) *Minimax Threshold*: The minimax threshold method uses a fixed threshold selected to produce minimax performance for the mean square error. The minimax uses a single threshold for all wavelet coefficients. It is defined in [89] as

$$\delta_{\min i \max} = \inf_\delta \sup_\mu \left\{ \frac{R_\delta(\mu)}{n^{-1} + \min(\mu^2, 1)} \right\} \tag{19}$$

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

*3) IWT:* The wavelet is associated with building a model for a signal, a system, or a process with a set of special signals. The wavelet transform provides a time–frequency representation of the signal. The wavelet transform uses a multiresolution technique by which frequencies are analyzed with different resolutions.

In a DWT, the scale factors between levels are usually chosen to be powers of two. For the DWT, the mother wavelet is defined as

$$\psi_{j,k}(t) = 2^{-j/2}\psi(2^{-j}t - k) \tag{20}$$

where $j, k \in Z$.

The DWT is given by

$$W_{j.k} = 2^{-j/2} \int_{-\infty}^{\infty} x(t)\psi(2^{-j}t - k)dt \tag{21}$$

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

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

$$x(t) = \sum_{j=1}^{\infty}\sum_{k\in Z} W_{j,k}\psi_{j,k}(t) + \sum_{j=1}^{\infty} c_{j,k}\phi_{j,k}(t) \tag{22}$$

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

$$c_{j,k} = 2^{-j/2} \int_{-\infty}^{\infty} x(t)\phi_{j,k}(2^{-j}t - k)dt. \tag{23}$$

The IWT reconstructs the signal from its coefficients.

### B. Change-Point-Detection Procedures

A 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 differ.

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

The multiple change points for the process are defined as [57]

$$\begin{aligned} X_1 &= (x_1, \ldots, x_{\tau_1}) \sim f_1 \\ X_2 &= (x_{\tau_1+1}, \ldots, x_{\tau_2}) \sim f_2 \\ &\ldots\ldots\ldots\ldots\ldots\ldots \\ X_\theta &= \left(x_{\tau_{\theta-1}+1}, \ldots, x_{\tau_\theta}\right) \sim f_\theta \end{aligned} \tag{24}$$

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

We first use a wavelet-denoising method to preprocess the data, and then, we use a sequential estimation scheme in [57] 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

$$T(a) = \inf\{n : S_n \geq a\} \tag{25}$$

where

$$S_n = \max_{1\leq k\leq n} \sum_{i=k+1}^{n} \log \frac{g(x_i)}{f(x_i)} \tag{26}$$

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

$$\hat{S}_n = \max_{1\leq k\leq n} \sum_{i=k+1}^{n} \log \frac{\hat{g}(x_i)}{\hat{f}(x_i)} \tag{27}$$

and the stopping rule described in [39] is defined by

$$\hat{T}(a) = \inf\{n : \hat{S}_n \geq a\}. \tag{28}$$

The call-log data are not a classical change-point model type. However, we use the method as described in [57] to detect trend changes, either linear or exponential, in each segment between change points. The computation of the cusum change-point detection will be more complicated if there are unknown parameters in segments [57].

We use a least-squares method for nonlinear trends. Whether a trend 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 a minimum weighted sum of squared residuals

$$\hat{S}_n = \min_k \left\{ \sum_{i=1}^{k} \sqrt{\hat{f}_i}(x_i - \hat{f}_i) + \sum_{i=k+1}^{n} \sqrt{\hat{g}_i}(x_i - \hat{g}_i) \right\} \tag{29}$$

and the stopping rule is defined as

$$\hat{T}(\alpha) = \inf\{n : p_n \leq \alpha\} \tag{30}$$

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

*2) Postestimation:* The detected change point must be estimated after a stopping rule has detected it. The change-point estimator that we use is based on the cusum stopping rule $\hat{T}$ and the minimum p-value

$$\hat{\tau} = \arg \min_{1\leq k<\hat{T}} p(k, \hat{T}, X) \tag{31}$$

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 points, we use ANOVA F-type tests. If the test is significant, we repeat steps 1–3 to search for the next change point; else, the test is a

false change point, and the search continues based on an initial sequence after the last significant change point.

For the fitting linear trend $E(x_i) = a + bt_i$, we use a standard least-squares estimate

$$\hat{b} = \frac{\sum_i (x_i - \overline{x})(t_i - \overline{t})}{\sum_i (t_i - \overline{t})^2} \qquad \hat{a} = \overline{x} - \hat{b}\overline{t} \tag{32}$$

and for the fitting exponential trend $E(x_i) = \exp(a + bt_i) - 1$, we use

$$\hat{b} = \frac{\sum_i \log(1 + x_i)(t_i - \overline{t})}{\sum_i (t_i - \overline{t})} \qquad \hat{a} = \overline{\log(1 + x)} - \hat{b}\overline{t} \tag{33}$$

for the initial approximation.

When the preliminary estimator of a change point is obtained, we perform a refinement of this estimator by least-squares fitting from the segment in the neighborhood of the preliminary estimator. If the change points are not significant for the chosen level $\alpha$, they are removed and the corresponding segments merged.

After the iterations end, all the change points are significant at the chosen level $\alpha$.

We select a Coiflets5 wavelet and the minimax threshold method to denoise the data by the principles described earlier and then apply the sequential change-point detection method in [39].

We use both simulation data and the real data from the data sets.

The simulation data sets are randomly generated based on $X = (X_1, X_2, \ldots, X_\theta)$, $X_i \sim N(\mu_i, \sigma_i^2)$ for $i = 1, 2, \ldots, \theta$.

### C. Unusual Consumption Event Detection

Change-point-detection methods do not deal with bursts of short width in time series. Bursts in time series are related to events such as attacks in networks. We propose the inhomogeneous Poisson model for detecting these bursts, which we term as unusual consumption events.

We assume that the number of calls follows an inhomogeneous Poisson process and the call duration follows an inhomogeneous exponential distribution.

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

$$N = \begin{bmatrix} n_{11} & n_{12} & \ldots & n_{1k} \\ n_{21} & n_{22} & \ldots & n_{2k} \\ \ldots & \ldots & \ldots & \ldots \\ n_{71} & n_{72} & \ldots & n_{7k} \end{bmatrix} \tag{34}$$

is the matrix of the number of calls on seven days of a week and

$$D = \begin{bmatrix} d_{11} & d_{12} & \ldots & d_{1k} \\ d_{21} & d_{22} & \ldots & d_{2k} \\ \ldots & \ldots & \ldots & \ldots \\ d_{71} & d_{72} & \ldots & d_{7k} \end{bmatrix} \tag{35}$$

is the matrix of the call duration on seven days of a week.

Then, the Poisson density function for day $i$ is given by

$$P_{N_i}(N_i = n_{ij}) = \frac{e^{-\lambda_i} \lambda_i^{n_{ij}}}{n_{ij}!} \tag{36}$$

where $\lambda_i$ is the rate (average) of the number of calls for day $i$.

By the properties of the Poisson distribution, the mean $= \lambda_i$, the variance var $= \lambda_i$, and the standard error $\sigma = \pm\sqrt{\lambda_i}$.

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

$$P_{D_i}(D_i = d_{ij}) = \frac{1}{\mu_i} e^{-\frac{d_{ij}}{\mu_i}} \tag{37}$$

where $\mu_i$ is the mean of the call duration for day $i$.

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

Now using maximum-likelihood estimates [86] to estimate $\lambda_i$ for day $i$, the cumulated probability distribution function is

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

$$= \frac{e^{-k\lambda_i} \lambda_i^{\sum_{j=1}^{k} n_{ij}}}{\prod_{j=1}^{k} n_{ij}!} \tag{38}$$

$$\ln P_{N_i} = -k\lambda_i + (\ln \lambda_i) \sum_{j=1}^{k} n_{ij} - \ln \left( \prod_{j=1}^{k} n_{ij} \right) \tag{39}$$

$$\frac{d(\ln P_{N_i})}{d\lambda_i} = -k + \frac{\sum_{j=1}^{k} n_{ij}}{\lambda_i} = 0 \tag{40}$$

$$\hat{\lambda}_i = \frac{\sum_{j=1}^{k} n_{ij}}{k}. \tag{41}$$

For $\mu_i$, the cumulated probability distribution function of the call duration is

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

$$= \frac{1}{\mu_i^k} e^{-\frac{1}{\mu_i} \sum_{j=1}^{k} d_{ij}} \tag{42}$$

$$\ln P_{D_i} = -k \ln \mu_i - \frac{1}{\mu_i} \sum_{j=1}^{k} d_{ij} \tag{43}$$

$$\frac{d(\ln P_{D_i})}{d\mu_i} = -\frac{k}{\mu_i} + \frac{1}{\mu_i^2} \sum_{j=1}^{k} d_{ij} = 0 \tag{44}$$

$$\hat{\mu}_i = \frac{\sum_{j=1}^{k} d_{ij}}{k}. \tag{45}$$

The maximum-likelihood estimates are used to estimate the average number of calls and call duration. Next, we consider the maximum average number of calls and call duration obtained for all weekday/weekend and week by week. Suppose that the $m$ week data are used to compute the rates of the number of calls and call duration for user $p$. Let $\hat{\lambda}_{d1}^p, \hat{\lambda}_{d2}^p, \ldots, \hat{\lambda}_{d7}^p$ be the

rate of the number of calls obtained for all weekday/weekend and $\hat{\lambda}^p_{w1}, \hat{\lambda}^p_{w2}, \ldots, \hat{\lambda}^p_{wm}$ be the rate of the call duration obtained week by week for $m$ weeks for user $p$, respectively. Let $\hat{\mu}^p_{d1}, \hat{\mu}^p_{d2}, \ldots, \hat{\mu}^p_{d7}$ be the mean of the call duration obtained for all weekday/weekend and $\hat{\mu}^p_{w1}, \hat{\mu}^p_{w2}, \ldots, \hat{\mu}^p_{wm}$ be the mean of the call duration obtained week by week for $m$ weeks for user $p$, respectively.

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

$$\hat{\lambda}^p_{\max} = \max\left(\hat{\lambda}^p_{d1}, \hat{\lambda}^p_{d2}, \ldots, \hat{\lambda}^p_{d7}, \hat{\lambda}^p_{w1}, \hat{\lambda}^p_{w2}, \ldots, \hat{\lambda}^p_{wm}\right) \quad (46)$$

$$\hat{\mu}^p_{\max} = \max\left(\hat{\mu}^p_{d1}, \hat{\mu}^p_{d2}, \ldots, \hat{\mu}^p_{d7}, \hat{\mu}^p_{w1}, \hat{\mu}^p_{w2}, \ldots, \hat{\mu}^p_{wm}\right) \quad (47)$$

where $\hat{\lambda}^p_{\max}$ and $\hat{\mu}^p_{\max}$ are the maximum-likelihood estimates of the 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. Our assumption is that the calling pattern could be different. Each caller has his/her own thresholds, and if the number of calls or call duration is greater than their usual thresholds for some day, we define that some event has occurred in that day.

To calculate the threshold of the number of calls for user $p$, $N^p_{\text{thres}}$, we define

$$N^p_{\text{thres}} = \hat{\lambda}^p_{\max} + \hat{\sigma}^p_{\max} \quad (48)$$

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

To calculate the threshold of the call duration for user $p$, $D^p_{\text{thres}}$, we define

$$D^p_{\text{thres}} = \hat{\mu}^p_{\max} + \hat{\delta}^p_{\max} \quad (49)$$

where $\hat{\mu}^p_{\max}$ and $\hat{\delta}^p_{\max}$ are the maximum mean of the 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

$$C = \langle (t_1, a_1, d_1, l_1), (t_2, a_2, d_2, l_2), \ldots (t_n, a_n, d_n, l_n) \rangle$$

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 $\langle s_i, r_i \rangle$, where $s_i$ is an actor who initiates a call at time $t_i$ and $r_i$ is an actor who receives the call. An unusual consumption event is defined as a subset $E \subset C$ of a tuple $E = \{(t_1, a_1, d_1, l_1), (t_2, a_2, d_2, l_2), \ldots, (t_m, a_m, d_m, l_m)\}$ such that either $\sum_{i=1}^{m} d_i > D_{\text{thres}}$ or $count(d_i) > N_{\text{thres}}$ defined as mentioned earlier in the time period $\Delta t = t_m - t_1$.

## VI. Pattern Recognition

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. In addition, customers can easily choose to be removed from a mailing or phone list if the advertisements are unwanted, thus eliminating unsolicited e-mails or phone calls. People may be interested in some advertisements for a period of time but will not want to receive those advertisements later. Ultimately, the customer comes to consider this traffic as spam and decides to opt out. We believe that current spam filters have great difficulty detecting this type of traffic. Note that several kinds of opt-ins exist. We consider opt-ins whom customers show a lot of interest for a short period of time and later have no interest but still keep getting unwanted e-mails or calls as *opt-in* bursts.

Another instance where researchers and developers can find pattern recognition useful occurs with presence awareness. Homeland security agents sometimes want to know when potential terrorists would like to communicate with their partners by modern telecommunication devices. With this information, agents can trace, scout, surveil, and detect potential terroristic attacks and evidence. We propose the use of a Bayesian inference model to compute the willingness level of people's communications with one another at a given time. Another example of willingness level of people's communications is a computer and telecommunication presence. The emergence of presence-aware communications allows people to quickly connect with others via the best choice of communication means, whether on the road, in meetings, or working from remote locations. Presence awareness also lets users know when others in their contact list are online. For those interested in studying presence awareness, presence information can include more user details, such as availability, location, activity, device capability, and other communication preferences. Researchers and developers can use presence to detect and convey willingness and ability to talk on the phone. Presence-enabled telephony services can reduce telephone traffic, as well as 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.

### A. Opt-in Detection

Opt-in burst detection is related to burst detection on data streams and to time series which are continuous data. However, the opt-in behavior resembles accumulated impulses and is not continuous. The approaches discussed in [93]–[101] to detect bursts are used for text, novel, and unusual data points or segments in time series that either have contents or are traffic data. *However, none of the previous work focuses on the specific problem that we study in this paper—opt-in bursts by studying the calling pattern based on call-detail records to detect opt-in bursts that reflect human activity*.

We define the opt-in bursts as dense sequences of 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

$$b_i(t) = \sum_{j=1}^{w_i} s_t \delta\left(t - t^i_j\right) \quad (50)$$

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

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

$$A_i = \frac{1}{w_i} \sum_{j=1}^{w_i} s_t \delta\left(t - t_j^i\right) \tag{51}$$

where $s_t$ is the value of an impulse at point $t$.

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

$$SW_k(t) = A_k rect\left(\frac{t - t_m}{\tau_k}\right) \tag{52}$$

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

*Definition of an Opt-in Burst:* A collection of call-log data can be represented as

$$C = \langle (t_1, a_1, d_1, l_1), (t_2, a_2, d_2, l_2), \ldots (t_n, a_n, d_n, l_n) \rangle$$

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 $\langle s_i, r_i \rangle$, where $s_i$ is an actor who initiates a call at time $t_i$ and $r_i$ is an actor who receives a call. An opt-in burst is defined as a subset $E \subset C$ of a tuple $E = \{(t_1, a_1, d_1, l_1), (t_2, a_2, d_2, l_2), \ldots, (t_m, a_m, d_m, l_m)\}$ such that $0 < count(d_i) < N_{\text{thres}}$ in the time period $\Delta t = t_m - t_1$, where $N_{\text{thres}}$ is a threshold which can be estimated from the historical data.

We process the sequence $S$ by an exponentially weighted moving average (EWMA) and then apply the dynamic-size sliding windows to detect opt-in bursts. The EWMA places more emphasis on the most recent data. Therefore, the EWMA would be more useful in dynamic systems.

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

$$\overline{s}_k = \frac{1}{M} \sum_{i=k-M+1}^{k} s_i \tag{53}$$

where $\bar{s}_k$ is the MA of $k$'s instance and $M$ is the number of the latest values. The EWMA can be derived from MA as

$$\overline{s}_k = (1-\alpha)s_k + \alpha(1-\alpha)s_{k-1} + \alpha^2(1-\alpha)s_{k-2} + \alpha^3 \overline{s}_{k-3} \tag{54}$$

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

### B. Willingness Level Inference

When callers want to make a call, they would like to know if the callee is in a mood to receive a call. In other words, callers would like to know when it is a good time to call the particular callees. We estimate the chance that a callee will accept a call based on the time of the day, call duration, and the location.

We propose a Bayesian inference model to compute a receiver's willingness level in a given time.

Let $X$ and $Y$ be two events. By the conditional probability rule [102], the probability of an event $X$ given $Y$ is

$$P(X|Y) = \frac{P(X,Y)}{P(Y)} \tag{55}$$

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

By the chain rule of conditional probability [102], we have

$$P(X,Y) = P(X|Y)P(Y). \tag{56}$$

Because the order of $X$ and $Y$ does not matter in (20), we have

$$P(Y,X) = P(Y|X)P(X). \tag{57}$$

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

$$P(X|Y)P(Y) = P(Y|X)P(X).$$

Thus, we have Bayes' theorem [102]

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)}. \tag{58}$$

In (58), $P(X|Y)$ is called the posterior probability, $P(Y|X)$ is referred to as the likelihood, and $P(X)$ is the prior probability.

Let $Y = (y_1, y_2, \ldots, y_n)$ depict $n$ measurements made on the sample from $n$ attributes $A_1, A_2, \ldots, A_n$, respectively. Assume that there are $m$ classes $x_1, x_2, \ldots, x_m$. Given an unknown data sample $X$, by Bayes' theorem

$$P(x_i|Y) = \frac{P(Y|x_i)P(x_i)}{P(Y)}. \tag{59}$$

$P(Y)$ is constant for all classes. $P(x_i) = n_i/N$, where $n_i$ is the number of training samples of class $x_i$, and $N$ is the total number of training samples.

If there are many attributes in the data set, it is extremely computationally expensive to compute $P(Y|x_i)$. In order to reduce the computation, we assume that the classes are conditionally independent. Thus

$$P(Y|x_i) = \prod_{j=1}^{n} P(y_j|x_i). \tag{60}$$

If $A_j$ is categorical, then $P(y_j|x_i) = n_{ij}/\sum_j n_{ij}$, where $n_{ij}$ is the number of training samples of class $x_i$ having the value $y_j$ for $A_j$ and $\sum_j n_i$ is the number of training samples belonging to $x_i$.

Let $X = (\text{incoming call} = \text{accept}, \text{incoming call} = \text{missed})$.

Let $Y = (T_i, D_j, Loc_l)$, where $T_i$ is the time interval, $i = 0, 1, 2, \ldots, 23$, (e.g., 0–1 o'clock); $D_j$ is a day, $j = 1, 2, \ldots, 7$, i.e., $D_1 = \text{Sunday}, D_2 = \text{Monday}, \ldots, D_7 = \text{Saturday}$; and $Loc_l = \text{location name}, l = 1, 2, \ldots, n$.

By Bayes' theorem, the willingness level to accept a call is

$$P(incoming = accept|T_i, D_j, Loc_l) = \frac{P(T_i, D_j, Loc_l|incoming = accept)P(incoming = accept)}{P(T_i, D_j, Loc_l)}. \tag{61}$$

## VII. Experimental Results and Discussions

Each of the approaches proposed earlier is validated here based on the data set described in the next section.

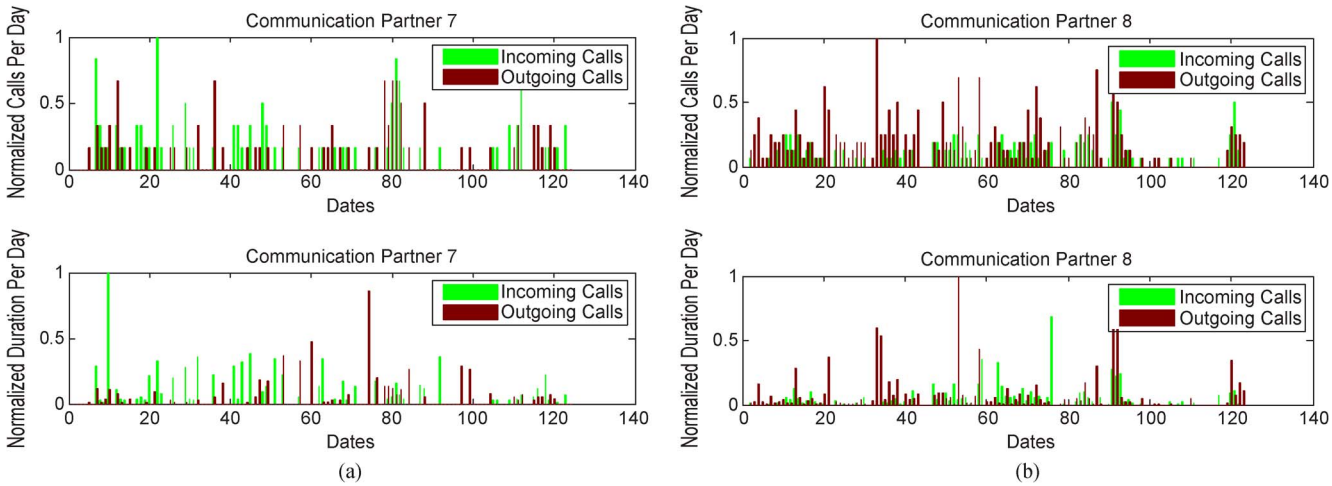


Fig. 3. (a) Socially close member of user12 by feedback from UNT data. (b) Socially close member of user70 by hand labeling from MIT data.

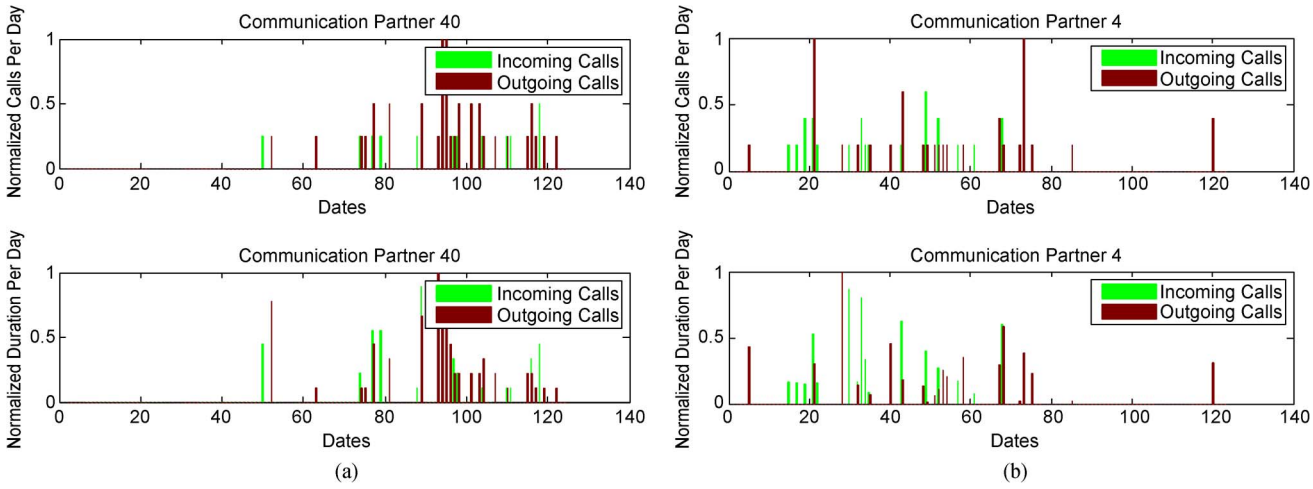


Fig. 4. (a) Socially near member of user12 by feedback from UNT data. (b) Socially near member of user70 by hand labeling from MIT data.

### A. *Real-Life Data Sets and Parameters*

*Real-life traffic profile*: In this paper, actual call logs are used for analysis. The call logs of 81 users which contain approximately 500 000 h of data on users' location, communication, and device usage behavior were collected for a period of eight months at the Massachusetts Institute of Technology (MIT) [103] by the Reality Mining Project group. Additionally, the call logs of 20 users were collected for a period of six months by the Network Security team at the University of North Texas (UNT). There is a total of around 5000 callers in the MIT and UNT data sets.

The Reality Mining Project group collected data on the mobile-phone usage of 81 users, including user ID (unique number representing a mobile-phone user), time of call, call direction (incoming or outgoing), incoming call description (missed or 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 to gather feedback from participating phone users about behavior patterns such as favorite hangout places; service providers; talk-time minutes; and phone users' friends, relatives, and parents. More information about the Reality Mining Project can be found in [103].

*Hand-Labeling*: To evaluate our methods' accuracy, we randomly choose 20 phone users from the MIT data set. We used our UNT data set as a pattern for hand labeling the MIT data. That is, the UNT data set contains direct user feedback that identifies socially related groups. The identified patterns were then used to hand label the MIT data and to validate our models and methods. Since the two groups are similarly composed of students, professors, and staff, our conjecture is that they will present similar behavior. We hand labeled the MIT communication members based on the number of calls, duration of calls in the period, history of call logs, location, and time of arrivals.

Figs. 3–5 show how we hand labeled the members in the MIT data set, where, in the left-hand side (a), the socially close, near, and far members were identified from the user's feedback in the UNT data. Corresponding members identified by hand labeling the MIT data appear in the right-hand side (b); the $x$-axis indicates the days, and the $y$-axis indicates the normalized number of calls and call duration, respectively.

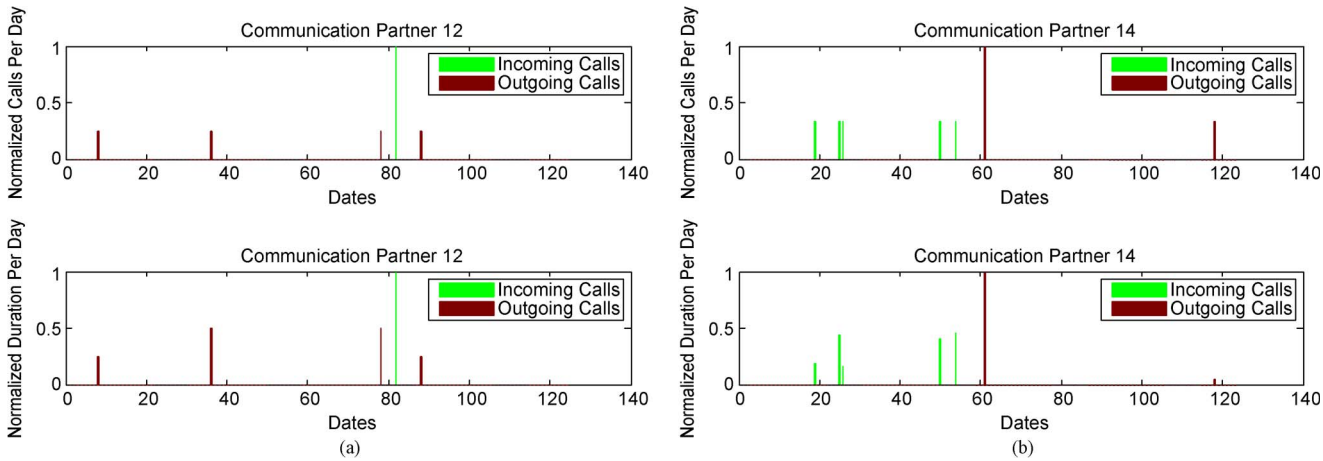


Fig. 5. (a) Socially far member of user12 by feedback from UNT data. (b) Socially far member of user70 by hand labeling from MIT data.

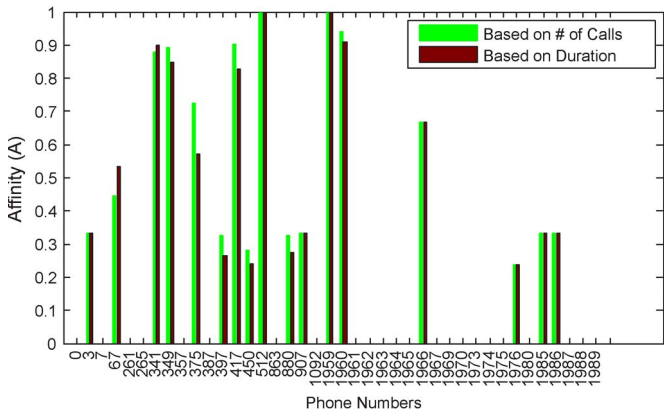


Fig. 6. Affinity values for phone user29.

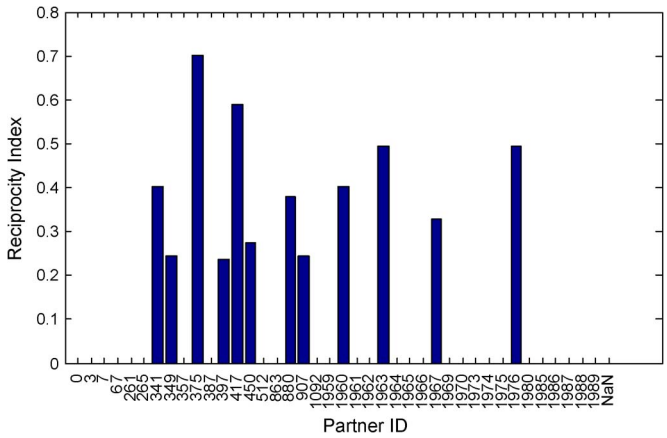


Fig. 7. Reciprocity indices for phone user29.

### B. Quantifying Social Groups and Reciprocity Index

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

In Fig. 6, the $x$-axis indicates the phone numbers that were used to communicate with user29 for four months. The $y$-axis indicates the affinity values based on both number of calls and call duration, respectively. From Fig. 6, we can see that user29 has seven socially close members, eight socially near members, and 24 socially far members during this four-month period. We divide the social-group members into work-time members and nonwork-time members. In general, during work time, the majority of our phone calls will be to talk with colleagues, bosses, secretaries, clients, and customers. We only occasionally speak to family members and friends, and during our nonwork time, we usually talk with family members and friends; we may speak with colleagues, clients, and customers only in special cases. Note that some people may be both our work-time colleagues and nonwork-time friends. Thus, the set of work-time members and the set of nonwork-time members may overlap. User29, a student, had one socially close member, two socially near members, and one socially far member in the work time and four socially close members, 11 socially near members, and 23 socially far members in the nonwork time. Note that, during the work time, user29 also has used office or public phones to speak to his colleagues. However, in this paper, we only use the cell phone call log 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 the work time.

Fig. 7 shows the reciprocity index results of user29 with communication partners, where the $x$-axis indicates the phone numbers and the $y$-axis indicates the reciprocity index values. User29 has 39 communication partners. For example, the reciprocity index is 0.72 for communication partner 375 and 0 for communication partner 7.

To find the relationships between the reciprocity index and specific time intervals of reciprocity, we calculated the reciprocity index for time intervals $t_j - t_i$ which is equal to $1, 2, \ldots, 24$ h, respectively.

Fig. 8 shows the reciprocity index for user39 with communication partner 316, where the $x$-axis indicates the time intervals in hours and the $y$-axis indicates the reciprocity index values. Fig. 8 shows a decreasing trend of the reciprocity index values when the time intervals increase.

Fig. 9 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 user39 with his partner 316, who was a frequent communication

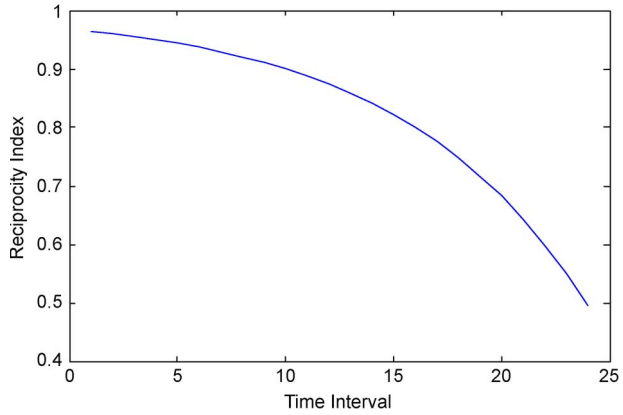


Fig. 8. Reciprocity index values for phone user39 with his partner 316.

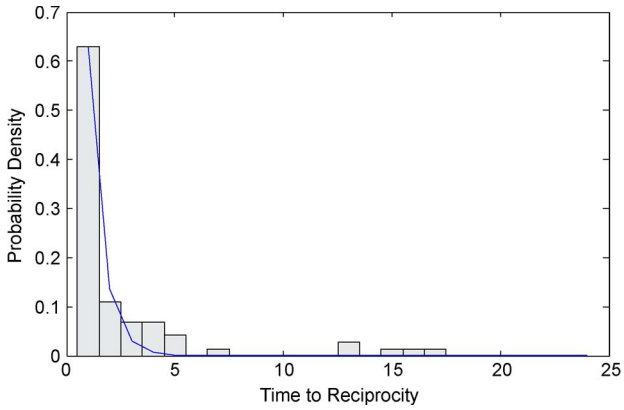


Fig. 9. Probabilities of reciprocity for phone user39 with his partner 316.

partner. From Fig. 9, we can see an exponential distribution for the reciprocity time, and most reciprocity time is within 1 h. Reciprocity distributions follow exponential trends for frequent communication partners. We also found that the cases in which the reciprocity time was greater than 10 h usually occurred when actor $a$ called actor $b$ close to evening sleeping time and actor $b$ called actor $a$ back on the next day.

By distribution fitting, we have probability density function

$$f(t) = 2.8e^{-1.5t} \tag{62}$$

for user39 with partner 316.

Table I describes the experimental results for 20 phone users for social groups. Our affinity model achieved performance with an accuracy of 94%.

We found that the fail or unsure cases occurred only when the number of calls was few, was either incoming or outgoing, and was consistent in each month of the four-month period that we studied. However, these kinds of cases seldom occurred in our experiments.

### C. Change-Point Detection

We evaluated the change-point detection first using synthetic data generated by simulation and then using real data.

Many workers have weeks when they are very busy (we refer to these as "busy weeks") and others where the work to be performed is less than their regular load (we refer to these as "easy weeks"). The expectation is that, during busy weeks, a worker makes and receives fewer calls and has shorter talk times than during usual weeks. During easy weeks, there is a tendency that a worker will make and receive more calls with a longer talk time than during usual weeks. To test this expectation, we randomly generated multiple simulation data sets of calls and call duration for 120 days. Figs. 10 and 11 are examples of the behavior that we observed.

TABLE I
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 |

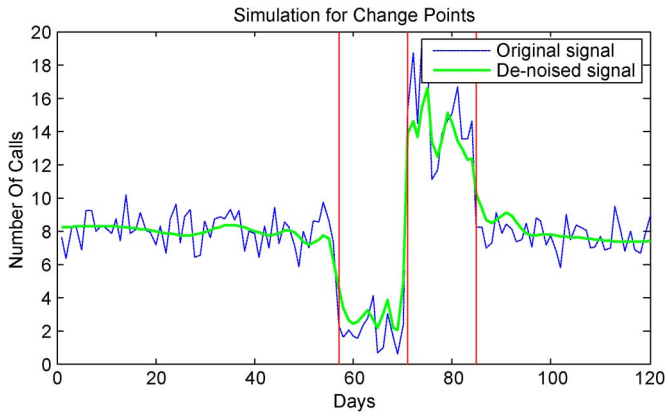


Fig. 10. Change points based on number of calls for simulation data.

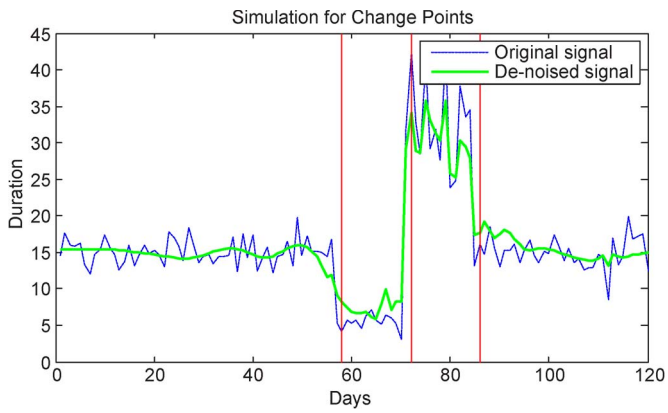


Fig. 11. Change points based on call duration for simulation data.

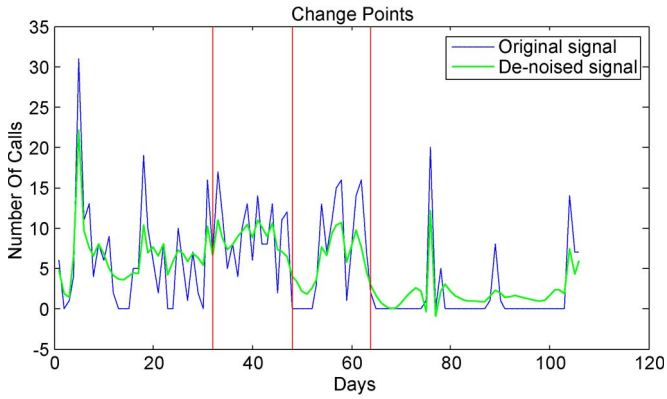


Fig. 12. Change points based on number of calls for user3.

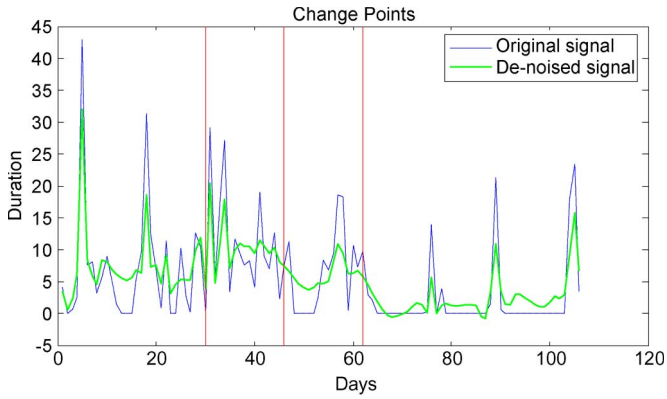


Fig. 13. Change points based on duration for user3.

Figs. 10 and 11 show the change points for the number of calls and call duration of data set1 and data set2, where the $x$-axis indicates the days and $y$-axis indicates the number of calls and call duration (in minutes), respectively. The blue dotted curve and green solid curve indicate original and denoised curves, respectively. The vertical lines indicate the change points. The three detected change points happened on the 56th, 60th, and 84th days. These days match the change points of the curves.

After multiple experiments with real data, we selected a Coiflets5 wavelet and a minimax threshold method by the principles described in Section 5 to denoise the data, and then, we applied the sequential change-point detection method as described in [40].

Figs. 12 and 13 show the change points for the number of calls and call duration of user3, 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, the original and denoised data. The vertical lines indicate the change points. There are three change points identified at the 32nd, 48th, and 64th days, which correspond to Friday, Sunday, and Sunday, respectively.

From the 1st day to the 32nd day, user3 visited New York City, the World Trade Center, and Harvard University. The average of the number of calls was eight, and the average of the duration was 7.5 min a day. Between the 32nd and 48th days, the user's activities were in local areas. The average

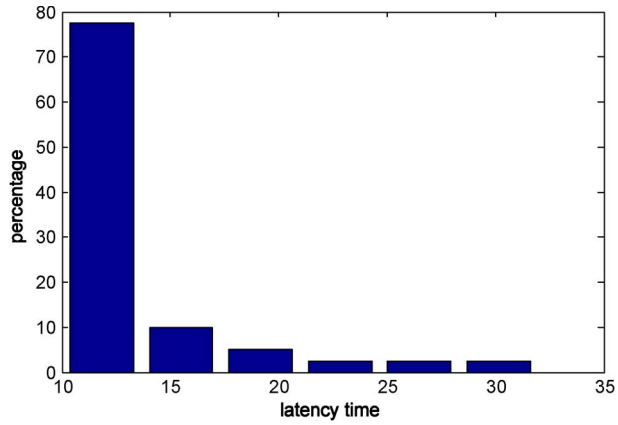


Fig. 14. Latency in the change-point detection for 20 users.

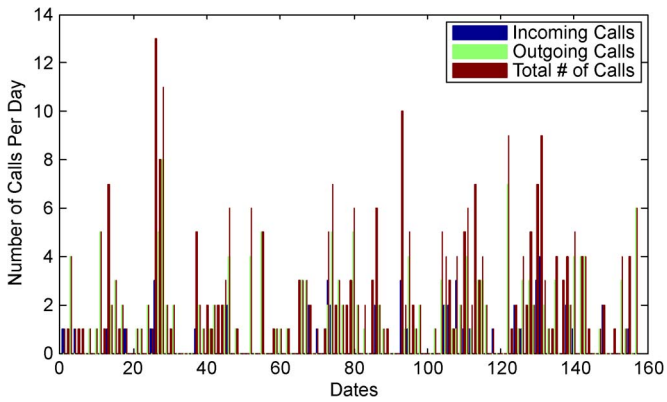


Fig. 15. Number of incoming, outgoing, and total calls per day for user74.

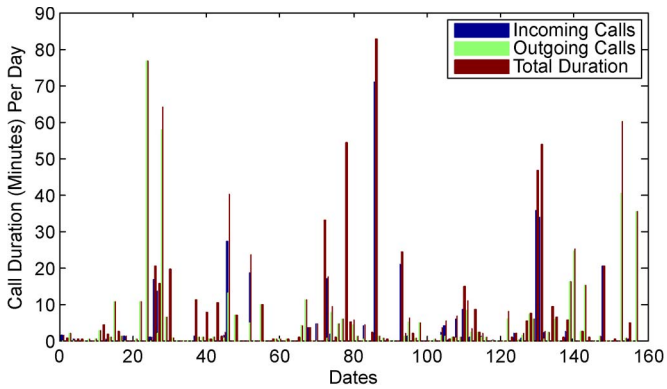


Fig. 16. Duration of incoming, outgoing, and total calls per day for user74.

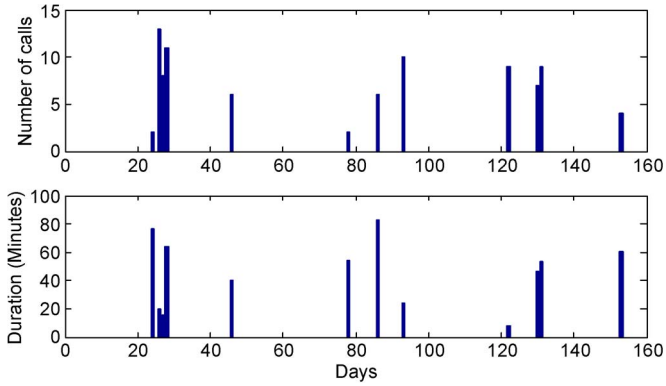


Fig. 17. There are events for these days for user74.

TABLE II
EVENT DATES AND LOCATIONS

| Users | Event | Days | # of contacts | # of calls the day | Duration (minutes) | Location | Note |
|---|---|---|---|---|---|---|---|
| User3 Threshold of # of calls a 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 long duration |
| | 2 | 18th | 3 | 19 | 31.3 | Visit Harvard Univ. | Both large # of calls and long 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 a day is 7 Threshold of duration a day is 36 (minutes) | 1 | 24th | 1 | 2 | 76.8 | Campus | Long 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 | Long duration |
| | 6 | 78th | 1 | 2 | 54.4 | At home | Long duration |
| | 7 | 86th | 2 | 6 | 82.8 | At home | Long 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 | Long duration |
| | 12 | 131st | 2 | 9 | 53.7 | Home (next day visit Providence RI) | Both large # of calls and long duration |
| | 13 | 153rd | 2 | 4 | 60 | Visit New York | Long duration |

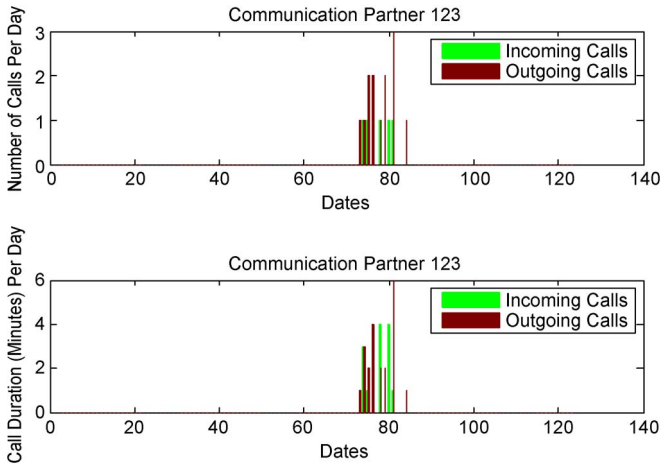


Fig. 18. Opt-in bursts for user1.

of the number of calls was 12, and the average of the duration is 10 min a day. From the 48th day to the 64th day, the user's activities were again within local areas. The average of the number of calls was nine, and the average of the duration is 6 min a day. From the 64th day to the 108th day, the user's activities were within local areas. The average of the number of calls was two, and the average of the duration was 2 min a day.

From the results, we found that most change-point days were associated with weekends, when people tended to have leisure time which, consequently, changed their calling patterns. Although a sophisticated technique was not needed to find change points over the weekends, it should be noted that our goal here is to show that our technique can identify change points based on call-detail records. The actual change point can, of course, represent other behaviors; for example, the change point may indicate whether the individual under observation is a potential threat to public security.

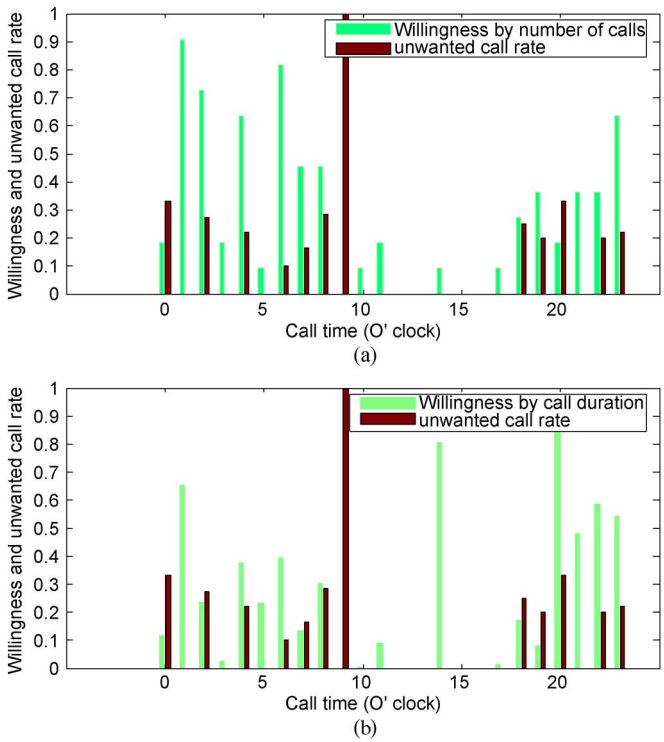


Fig. 19. Willingness level on Sundays. (a) Willingness level compared to unwanted-call rate (rejected/missed) by the number of calls. (b) Computed willingness level based on talk time.

To evaluate the performance of our change-point detection method, we conducted experiments on 20 phone users randomly selected from the call logs of 81 phone users. The significant level and latency are two important measurements when considering the accuracy in change-point detection. The lower the significant level, the more sensitive the strategy. This can help to avoid false positives. For all data sets, we used a significant level of 0.01.

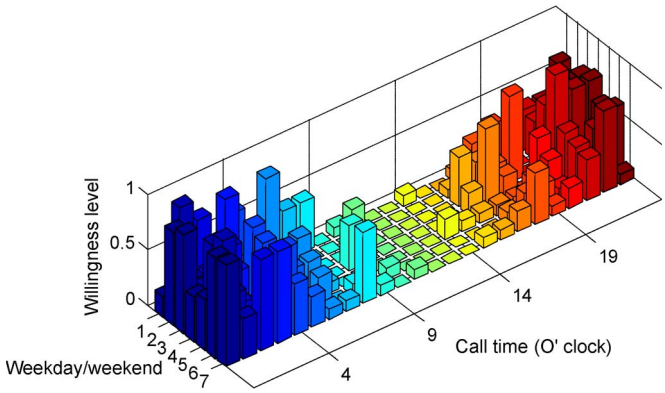


Fig. 20. Willingness level during 24 h from Sunday to Saturday.

The shorter the latency, the better the indication of the strategy's accuracy. Therefore, a minimization of the latency is desired. In this paper, we chose a minimum of ten data points to start the detection of a new change point. We identified a total of 81 change points for 20 users. 78% of the change points were identified using about 12 data points and 10% using about 15 data points as shown in Fig. 14. In only a few cases were the change points identified later than the 20th points (see Fig. 14.) This result is a good indication of the strategy's accuracy.

### D. Unusual Consumption Event Detection

Figs. 15 and 16 show the number of calls and call duration for user74, while Fig. 17 shows the event days. From Fig. 17, we may see that user74 had 12 event days (the 24th, 26th, 27th, 28th, 46th, 78nd, 86th, 93rd, 122nd, 130th, 131st, and 153rd days of 160 days). During these event days, there occurred either a large number of calls or a long call duration. On six of these days, user74 had traveled to New York City and Providence, RI.

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

For example, in Table II, there were 17 calls, which was greater than 15 (the threshold of the number of calls) on the 33rd day for user3. We define that there was some event in that day. For user74 on the 24th day, although there were only two calls, the call duration was 76.8 min, which was much longer than 18 min (the threshold of the call duration), and there was some event in that day.

### E. Opt-in Detection

Fig. 18 shows the opt-in bursts for user1 with 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. User1 first called communication partner123.

For 20 phone users with a total of 1863 communication partners, we found that 1.4% of them are opt-in bursts.

TABLE III
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 |

### F. Willingness Level Inference

We calculated the willingness level to receive calls and the corresponding unwanted-call rate for users for 1-h intervals from 0 to 23 o'clock from Sunday to Saturday.

In Fig. 19, the $x$-axis indicates the calling time for incoming as well as outgoing calls for 24 h on Sunday. The $y$-axis indicates the willingness level for a second-year graduate-student user. In this analysis, missed calls were considered as unwanted calls which were compared with the willingness level. When the willingness level was low, there were more missed calls.

Fig. 19(b) describes the willingness level calculated based on total talk time. From Fig. 19(a) and (b), we can see that, when the user is more willing to receive calls, the number of missed calls decreases. The receiver missing calls means that these are unwanted calls at that specific time. For example, in Fig. 19(a), the willingness level is 0.7 (70%) and the corresponding unwanted-call rate is 0.28 (28%) between 2 and 3 o'clock. As one more example, the willingness level is 0.2 (20%) and the corresponding unwanted-call rate is 0.33 (33%) between 0 and 1 o'clock.

Fig. 20 shows the willingness level of this graduate-student user based on the number of calls received from Sunday to Saturday. Here, the $x$-axis represents the time of the day, and the $y$-axis represents the seven days of a week. The first unit on the $y$-axis represents Sunday, and the last unit represents Saturday.

We validated the user's willingness level with respect to the number of missed or rejected calls. We measured the accuracy by the unwanted-call rate over the range of willingness levels. The unwanted-call rate is a ratio of the number of missed calls to the total number of calls within a given time period. The assumption was that a missed call was an unwanted call.

Table III and Fig. 21 describe the experimental results for ten phone users. In Table III, our results show that our model achieves good performance. For example, when the willingness level was 0%–30%, the average unwanted rate was 43.32% with a standard error of 1.79%. The mean unwanted rate was 4.84% for the willingness level of 71%–100%. We found that the higher the willingness level, the lower the unwanted rate, and vice versa.

## VIII. CONCLUSION

In this paper, we proposed a socioscope model for social-network and human-behavior analysis based on mobile-phone

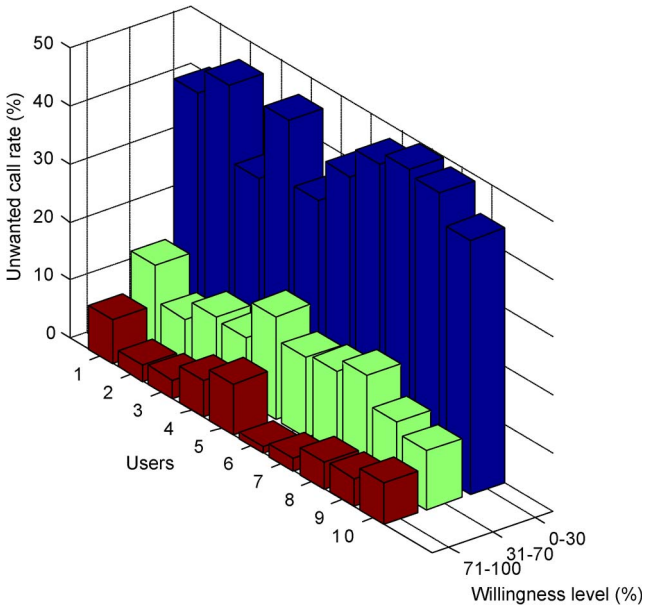


Fig. 21. Willingness level (in percent) versus unwanted-call rate (in percent) for ten users.

call-detail records. Because of the diversities and complexities of human social behavior, one technique cannot detect all the different features of human social behaviors. Thus, we used multiple probability and statistical methods, including *affinity, wavelet denoising, sequential detection, inhomogeneous Poisson and inhomogeneous exponential model, and Bayesian inference* for quantifying social groups, relationships, and communication patterns and for detecting human-behavior changes.

As a result of this analysis, we propose a new *reciprocity index* to measure the level of reciprocity between users and their communication partners.

We may quantify relationships for a short-term period, e.g., a month, or a long-term period, e.g., a year or more, using our model by adjusting the parameters. Errors appear when the number of calls is small. However, these kinds of cases seldom occurred in our experiments.

This paper is useful for homeland security and for detecting unwanted calls, e.g., spam, telecommunication presence, and marketing. The experimental results show that our model achieves high accuracy. In our future work, we plan to investigate more information such as semantics, users' relation from other sources, other events, social-network dynamics, and evolution. Currently, we are working on applying our affinity model to quantify social relationships on tweeter followers of Twitter data. In addition, we also investigate some other strategies on Twitter data. We apply two methods for the classification of different social-network users such as leaders (e.g., news groups), lurkers, spammers, and close associates. The first method is a two-stage process with a fuzzy-set theoretic approach to evaluate the strengths of network links (user–user relationships), followed by a simple linear classifier to separate the user classes. The second method performs user classification by matching their short-term tweet patterns with the generic tweet patterns of the prototype users of different classes for handling the situation of limited availability of user data for learning network link strengths.

## ACKNOWLEDGMENT

The authors would like to thank N. Eagle and the Massachusetts Institute of Technology for providing the call logs of the Reality Mining data set. The authors would also like to thank the reviewers and editors for their valuable comments in improving the quality of this paper.

## REFERENCES

[1] S. Wasserman and K. Faust, *Social Network Analysis: Methods and Applications*. Cambridge, U.K.: Cambridge Univ. Press, 1994, pp. 505–555.
[2] Ogatha, "Computer supported social networking for augmenting cooperation," in *Computer Supported Cooperative Work*, vol. 10. Norwell, MA: Kluwer, 2001, pp. 189–209.
[3] M. E. J. Newman, "Modularity and community structure in networks," *Proc. Nat. Acad. Sci.*, vol. 103, no. 23, pp. 8577–8582, Jun. 2006.
[4] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener, "Graph structure in the web," *Comput. Netw.*, vol. 33, no. 2, pp. 309–320, Jun. 2000.
[5] P. Doreian, V. Batageli, and A. Ferligoj, *Generalized Blockmodeling*, M. Granovetter, Ed. Cambridge, U.K.: Cambridge Univ. Press, 2005.
[6] G. Flake, S. Lawrence, C. L. Giles, and F. Coetzee, "Self-organization and identification of web communities," *IEEE Comput.*, vol. 35, no. 3, pp. 66–70, Mar. 2002.
[7] J. Hopcroft, O. Khan, B. Kulis, and B. Selman, "Natural communities in large linked networks," in *Proc. 9th Int. Conf. Knowl. Discov. Data Mining*, 2003, pp. 541–546.
[8] M. E. J. Newman, "Detecting community structure in networks," *Eur. Phys. J.*, vol. B 38, no. 2, pp. 321–330, Mar. 2004.
[9] C. Borgs, J. Chayes, M. Mahdian, and A. Saberi, "Exploring the community structure of newsgroups," in *Proc. 10th ACM Int. Conf. Knowled. Discov. Data Mining*, 2004, pp. 783–787.
[10] P. Holme and M. Newman, "Nonequilibrium phase transition in the coevolution of networks and opinions," *Phys. Rev. E*, vol. 74, no. 5, p. 056 108, Mar. 2006.
[11] P. Sarkar and A. Moore, "Dynamic social network analysis using latent space models," *Proc. SIGKDD Explorations: Special Edition Link Mining*, vol. 7, no. 2, pp. 31–40, Dec. 2005.
[12] L. Backstrom, D. Huttenlocher, and J. Kleinberg, "Group formation in large social networks: Membership, growth, and evolution," in *Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2006, pp. 44–54.
[13] Kossinets and D. Watts, "Empirical analysis of an evolving social network," *Science*, vol. 311, no. 5757, pp. 88–90, Jan. 2006.
[14] X. Wang and A. McCallum, "Topics over time: 'A non-Markov continuous-time model of topical trends," in *Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2006, pp. 424–433.
[15] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins, "Structure and evolution of blogspace," *Commun. ACM*, vol. 47, no. 12, pp. 35–39, Dec. 2004.
[16] R. Kumar, J. Novak, and A. Tomkins, "Structure and evolution of online social networks," in *Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2006, pp. 611–617.
[17] A. A. Nanavati, S. Gurumurthy, G. Das, D. Chakraborty, K. Dasgupta, S. Mukherjea, and A. Joshi, "On the structural properties of massive telecom graphs: Findings and implications," in *Proc. 15th ACM CIKM Conf. Inf. Knowl. Manage.*, 2006, pp. 435–444.
[18] J. P. Onnela, J. Saramaki, J. Hyvonen, G. Szabo, D. Lazer, K. Kaski, J. Kertesz, and A. L. Barabasi, "Structure and tie strengths in mobile communication networks," *Proc. Nat. Acad. Sci. U.S.A.*, vol. 104, no. 18, pp. 7332–7336, May 1, 2007.
[19] M. Kurucz, A. Benczur, K. Csalogany, and L. Lukacs, "Spectral clustering in telephone call graphs," in *Proc. Joint 9th WEBKDD and 1st SNA-KDD Workshop*, 2007, pp. 82–91.
[20] W. Teng and M. Chou, "Mining communities of acquainted mobile users on call detail records," in *Proc. 22nd Annu. ACM Symp. Appl. Comput.*, 2007, pp. 957–958.
[21] G. Palla, A. Barabasi, and T. Vicsek, "Quantifying social group evolution," *Nature*, vol. 446, no. 7136, pp. 664–667, Apr. 2007.
[22] N. Eagle, A. Pentland, and D. Lazer, "Inferring social network structure using mobile phone data," *Proc. Nat. Acad. Sci. (PNAS)*, vol. 106, no. 36, pp. 15 274–15 278, Sep. 2009.

[23] C. A. Hidalgo and C. Rodriguez-Sickert, "The dynamics of a mobile phone network," *Phys. A*, vol. 387, no. 12, pp. 3017–3024, May 2008.

[24] K. Dasgupta, R. Singh, B. Viswanathan, D. Chakraborty, S. Mukherjea, and A. Nanavati, "Social ties and their relevance to churn in mobile telecom networks," in *Proc. 11th ACM Int. Conf. Extending Database Technol.: Adv. Database Technol.*, 2008, pp. 668–677.

[25] J. Candia, M. C. Gonzalezm, P. Wang, T. Schoenharl, G. Madey, and A. Barabasi, "Uncovering individual and collective human dynamics from mobile phone records," *J. Phys. A, Math. Theoretical*, vol. 41, no. 22, p. 224 015, Jun. 2008.

[26] A. Pentland, "Collective intelligence," *IEEE Comput. Intell. Mag.*, vol. 1, no. 3, pp. 9–12, Aug. 2006.

[27] A. Pentland, "Automatic mapping and modeling of human networks," *Phys. A, Stat. Mech. Appl.*, vol. 378, no. 1, pp. 59–67, May 2007.

[28] N. Eagle, "Behavioral inference across cultures: Using telephones as a cultural lens," *IEEE Intell. Syst.*, vol. 23, no. 4, pp. 62–64, Jul./Aug. 2008.

[29] N. Eagle and A. Pentland, "Eigenbehaviors: Identifying structure in routine," *Behav. Ecol. Sociobiol.*, vol. 63, no. 7, pp. 1057–1066, May 2009.

[30] N. Eagle, Y. de Montjoye, and L. Bettencourt, "Community computing: Comparisons between rural and urban societies using mobile phone data," in *Proc. IEEE Int. Conf. Social Comput.*, 2009, pp. 144–150.

[31] N. Eagle, "Engineering a Common Good," in *Proc. IEEE Engaging Data Conf.*, Boston, MA, Oct. 12–13, 2009, pp. 1–3.

[32] N. Eagle, "txteagle: Mobile crowdsourcing," *Int., Des. Global Develop.*, vol. 5623, *LNCS*, pp. 447–456, 2009.

[33] B. Gallagher, H. Tong, T. Eliassi-Rad, and C. Faloutsos, "Using ghost edges for classification in sparsely labeled networks," in *Proc. 14th ACM SIGKDD Conf.*, 2008, pp. 256–264.

[34] P. Koutsourelakis, "Unsupervised group discovery and link prediction in relational datasets: A nonparametric Bayesian approach," Lawrence Livermore Nat. Lab. (LLNL), Livermore, CA, Tech. Rep. UCRL-TR-230743, 2007.

[35] I. Carreras, D. Miorandi, G. Canright, and K. E. Monsen, "Eigenvector centrality in highly partitioned mobile networks: Principles and applications," *Adv. Biol. Inspired Inf. Syst.*, vol. 69, pp. 123–145, 2007.

[36] L. Wang, Y. Jia, and W. Han, "Instant message clustering based on extended vector space model," *Adv. Comput. Intell.*, vol. 4683, *Springer-Verlag Lecture Notes Computer Science*, pp. 435–443, 2007.

[37] A. Chaintreau *et al.*, "The diameter of opportunistic mobile networks," in *Proc. ACM CoNEXT Conf.*, New York, 2007.

[38] P. Hui, E. Yoneki, S. Y. Chan, and J. Crowcroft, "Distributed community detection in delay tolerant networks," in *Proc. MobiArch*, Kyoto, Japan, 2007, ACM Press.

[39] N. Eagle and A. Pentland, "Reality mining: Sensing complex social systems," *Pers. Ubiquitous Comput.*, vol. 10, no. 4, pp. 255–268, May 2006.

[40] J. Lawrence, T. R. Payne, and D. De Roure, "Co-presence communities: Using pervasive computing to support weak social networks," in *Proc. WETICE*, 2006, pp. 149–156, IEEE Press.

[41] N. Eagle, "Machine perception and learning of complex social systems," Ph.D. dissertation, Program Media Arts Sci., Massachusetts Inst. Technol., Cambridge, MA, Jun., 2005.

[42] C. Kobashikawa, Y. Hatakeyama, F. Dong, and K. Hirota, "Fuzzy algorithm for group decision making with participants having finite discriminating abilities," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 39, no. 1, pp. 86–95, Jan. 2009.

[43] F. Carvalho and Y. Lechevallier, "Dynamic clustering of interval-valued data based on adaptive quadratic distances," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 39, no. 6, pp. 1295–1306, Nov. 2009.

[44] R. R. Yager, "Concept representation and database structures in fuzzy social relational networks," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 40, no. 2, pp. 413–419, Mar. 2010.

[45] U. Maulik, S. Bandyopadhyay, and I. Saha, "Integrating clustering and supervised learning for categorical data analysis," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 40, no. 4, pp. 664–675, Jul. 2010.

[46] M. Fannes and P. Spincemaile, "The mutual affinity of random measures," *Periodica Math. Hungarica*, vol. 47, no. 1/2, pp. 51–71, Sep. 2003.

[47] A. W. Gouldner, "The norm of reciprocity: A preliminary statement," *Amer. Sociol. Rev.*, vol. 25, no. 2, pp. 161–178, 1960.

[48] L. Katz and J. Powell, "Measurement of the tendency toward reciprocation of choice," *Sociometry*, vol. 18, no. 4, pp. 403–409, Nov. 1955.

[49] L. Katz and T. Wilson, "The variance of the number of mutual choices in sociometry," *Psychometrika*, vol. 21, no. 3, pp. 299–304, Sep. 1956.

[50] M. Schnegg, "Reciprocity and the emergence of power laws in social networks," *Int. J. Mod. Phys. C*, vol. 17, no. 8, pp. 1067–1076, 2006.

[51] D. Garlaschelli and M. I. Loffredo, "Patterns of link reciprocity in directed networks," *Phys. Rev. Lett.*, vol. 93, no. 26, p. 268 701, Dec. 2004.

[52] G. Zamora-López, V. Zlatić, C. Zhou, H. Štefančić, and J. Kurths, "Reciprocity of networks with degree correlations and arbitrary degree sequences," *Phys. Rev. E*, vol. 77, no. 1, p. 016 106, Jan. 2008.

[53] L. M. Floría, C. Gracia-Lázaro, J. Gómez-Gardeñes, and Y. Moreno, "Social network reciprocity as a phase transition in evolutionary cooperation," *Phys. Rev. E*, vol. 79, no. 2, p. 026 106, 2009.

[54] B. Hogan and D. Fisher, "Email reciprocity and personal communication bias," in *Proc. NetSci*, Bloomington, IN, May 22–25, 2006.

[55] M. Baron, "Nonparametric adaptive change-point estimation and on-line detection," *Sequential Anal.*, vol. 19, no. 1/2, pp. 1–23, 2000.

[56] M. Baron and N. Granott, "Consistent estimation of early and frequent change points," in *Proc. Shoresh Conf.*, Y. Haitovsky, H. R. Lerche, and Y. Ritov, Eds., New York, 2003, pp. 181–192, Foundations of Statistical Inference, Springer.

[57] J. W. Cangussu and M. Baron, "Automatic identification of change points for the system testing process," in *Proc. 30th Annu. IEEE Int. COMPSAC*, Chicago, IL, Sep. 18–21, 2006, vol. 1, pp. 377–384.

[58] A. E. Raftery and V. E. Akman, "Bayesian analysis of a Poisson process with a change-point," *Biometrika*, vol. 73, no. 1, pp. 85–89, 1986.

[59] T. Y. Yang and L. Kuo, "Bayesian binary segmentation procedure for a Poisson process with multiple change points," *J. Comput. Graphical Statist.*, vol. 10, no. 4, pp. 772–785, Dec. 2001.

[60] Y. Ritov, A. Raz, and H. Bergman, "Detection of onset of neuronal activity by allowing for heterogeneity in the change points," *J. Neuroscience Methods*, vol. 122, no. 1, pp. 25–42, Dec. 2002.

[61] B. P. Carlin, A. E. Gelfand, and A. F. M. Smith, "Hierarchical Bayesian analysis of change point problems," *Appl. Stat.*, vol. 41, no. 2, pp. 389–405, 1992.

[62] R. Lund and J. Reeves, "Detection of undocumented change points: A revision of the two-phase regression model," *J. Clim.*, vol. 15, no. 17, pp. 2547–2554, Sep. 2002.

[63] D. M. Hawkins, "Testing a sequence of observations for a shift in location," *J. Amer. Stat. Assoc.*, vol. 72, no. 357, pp. 180–186, Mar. 1977.

[64] K. J. Worsley, "On the likelihood ratio test for a shift in location of normal populations," *J. Amer. Stat. Assoc.*, vol. 74, no. 366, pp. 365–367, Jun. 1979.

[65] J. Chen and A. K. Gupta, "Testing and locating change points with application to stock prices," *J. Amer. Stat. Assoc.*, vol. 92, no. 438, pp. 739–747, Jun. 1997.

[66] T. D. Johnson, R. M. Elashoff, and S. J. Harkema, "A Bayesian change-point analysis of electromyographic data: Detecting muscle activation patterns and associated applications," *Biostatistics*, vol. 4, no. 1, pp. 143–164, Jan. 2003.

[67] P. Chu and X. Zhao, "Bayesian change-point analysis of tropical cyclone activity: The central North Pacific case," *J. Clim.*, vol. 17, no. 24, pp. 4893–4901, Dec. 2004.

[68] P. Fearnhead, "Exact and efficient Bayesian inference for multiple change point problems," *Stat. Comput.*, vol. 16, no. 2, pp. 203–213, Jun. 2006.

[69] K. Yamanishi and J. Takeuchi, "A unifying framework for detecting outliers and change points from time series," *IEEE Trans. Knowl. Data Eng.*, vol. 18, no. 4, pp. 482–492, Apr. 2006.

[70] C. Erdman and J. W. Emerson, "A fast Bayesian change point analysis for the segmentation of micro-array data," *Bioinformatics*, vol. 24, no. 19, pp. 2143–2148, Oct. 2008.

[71] T. Brants and F. Chen, "A system for new event detection," in *Proc. Int. ACM SIGIR Conf.*, 2003, pp. 330–337.

[72] F. Chen, A. Farahat, and T. Brants, "Story link detection and new event detection are asymmetric," in *Proc. HLT-NAACL*, 2003, vol. 2, pp. 13–15.

[73] G. Kumaran and J. Allan, "Text classification and named entities for new event detection," in *Proc. Int. ACM SIGIR Conf.*, 2004, pp. 297–304.

[74] X. Li and B. W. Croft, "Novelty detection based on sentence level patterns," in *Proc. ACM CIKM*, 2005, pp. 744–751.

[75] Z. Li, B. Wang, M. Li, and W.-Y. Ma, "A probabilistic model for retrospective news event detection," in *Proc. Int. SIGIR Conf.*, 2005, pp. 106–113.

[76] Q. Zhao and P. Mitra, "Event detection and visualization for social text streams," in *Proc. ICWSM*, Boulder, CO, 2007.

[77] Q. He, K. Chang, and E. P. Lim, "A model for anticipatory event detection," in *Proc. 25th Int. Conf. Conceptual Model. (ER)*, vol. 4215, *Springer LNCS*, 2006, pp. 168–181.

[78] Q. He, K. Chang, and E. P. Lim, "Anticipatory event detection via sentence classification," in *Proc. IEEE Int. Conf. Syst., Man, Cybern.*, 2006, pp. 1143–1148.
[79] G. Luo, C. Tang, and P. Yu, "Resource-adaptive real-time new event detection," in *Proc. Int. ACM SIGMOD*, 2007, pp. 497–508.
[80] V. Guralnik and J. Srivastava, "Event detection from time series data," in *Proc. 5th ACM SIGKDD*, 1999, pp. 33–42.
[81] J. Kleinberg, "Bursty and hierarchical structure in streams," in *Proc. 8th ACM SIGKDD*, 2002, pp. 91–101.
[82] E. Keogh, S. Lonardi, and B. Chiu, "Finding surprising patterns in a time series database in linear time and space," in *Proc. 8th ACM SIGKDD*, 2002, pp. 550–556.
[83] M. Salmenkivi and H. Mannila, "Using Markov chain Monte Carlo and dynamic programming for event sequence data," *Knowl. Inf. Syst.*, vol. 7, no. 3, pp. 267–288, Mar. 2005.
[84] A. Ihler, J. Hutchins, and P. Smyth, "Adaptive event detection with time-varying Poisson processes," in *Proc. ACM SIGKDD*, 2006, pp. 207–216.
[85] C. C. Yang, X. Shi, and C. P. Wei, "Discovering event evolution graphs from news corpora," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 39, no. 4, pp. 850–863, Jul. 2009.
[86] J. W. Harris and H. Stocker, *Handbook of Mathematics and Computational Science*. New York: Springer-Verlag, 1998, p. 824.
[87] S. Bregni, R. Cioffi, and M. Decina, "An empirical study on statistical properties of GSM telephone call arrivals," in *Proc. IEEE Global Telecommun. Conf.*, 2006, pp. 1–5.
[88] H. Zhang and R. Dantu, "Quantifying the presence for phone users," in *Proc. 5th IEEE Consum. Commun. Netw. Conf.*, 2008, pp. 883–887.
[89] D. Donoho and I. Johnstone, "Ideal spatial adaptation by wavelet shrinkage," *Biometrica*, vol. 81, no. 3, pp. 425–455, Sep. 1994.
[90] D. Donoho and I. Johnstone, "Adapting to unknown smoothing via wavelet shrinkage," *J. Amer. Stat. Assoc.*, vol. 90, no. 432, pp. 1200–1224, Dec. 1995.
[91] C. Stern, "Estimation of the mean of multivariate normal distribution," *Ann. Statist.*, vol. 9, no. 6, pp. 1135–1151, Nov. 1981.
[92] I. Johnstone and B. W. Silverman, "Wavelet thresholding estimators for data with correlated noise," *J. Roy. Stat. Soc., Ser. B*, vol. 59, no. 2, pp. 319–351, 1997.
[93] T. Fujiki, T. Nanno, Y. Suzuki, and M. Okumura, "Identification of bursts in a document stream," in *Proc. 1st Int. Workshop Knowl. Discovery Data Streams*, 2004, pp. 55–64.
[94] M. Vlachos, C. Meek, Z. Vagena, and D. Gunopulos, "Identifying similarities, periodicities and bursts for online search queries," in *Proc. ACM SIGMOD*, 2004, pp. 131–142.
[95] Y. Zhu and D. Shasha, "Efficient elastic burst detection in data streams," in *Proc. 9th ACM SIGKDD*, 2003, pp. 336–345.
[96] S. Qin, W. Qian, and A. Zhou, "Adaptively detecting aggregation bursts in data streams," in *Proc. 10th Int. Conf. Database Syst. Adv. Appl.*, vol. 3453, *LNCS*, 2005, pp. 435–446.
[97] T. Chen, Y. Wang, B. Fang, and J. Zheng, "Detecting lasting and abrupt bursts in data streams using two-layered wavelet tree," in *Proc. Adv. Int. Conf. Telecommun.*, 2006, p. 30.
[98] M. Wang, T. M. Madhyastha, N. H. Chan, S. Papadimitriou, and C. Faloutsos, "Data mining meets performance evaluation: Fast algorithms for modeling bursty traffic," in *Proc. 18th Int. Conf. Data Eng.*, 2002, pp. 507–516.
[99] G. P. C. Fung, J. X. Yu, P. S. Yu, and H. Lu, "Parameter free bursty events detection in text streams," in *Proc. 31st Int. Conf. Very Large Data Bases*, 2005, pp. 181–192.
[100] Z. Yuan, Y. Jia, and S. Yang, "Online burst detection over high speed short text streams," in *Proc. Int. Conf. Comput. Sci.*, 2007, pp. 717–725.
[101] M. Vlachos, K. Wu, S. Chen, and P. Yu, "Fast burst correlation of financial data," in *Proc. 9th Eur. Conf. Princ. Pract. Knowl. Discovery Databases*, 2005, pp. 368–379.
[102] N. Nilsson, *Artificial Intelligence, A New Synthesis*, 1st ed. San Fransisco, CA: Morgan Kaufmann, 1998, pp. 317–336.
[103] Massachusetts Institute of Technology, Reality Mining, 2009. [Online]. Available: http://reality.media.mit.edu/

**Huiqi Zhang** received the B.E. degree from the University of Science and Technology-Beijing, Beijing, China, the M.E. degree from the China Academy of Railway Sciences, Beijing, the M.S. degrees in mathematics and engineering from South Dakota State University, Brookings, and the Ph.D. degree in computer science from the University of North Texas, Denton.

He is currently with the Department of Computer Science and Engineering, University of North Texas. His research interests include information security, computer networks, artificial intelligence, data mining, and social computing.

**Ram Dantu** (M'09) received the B.S. degree from Madras Institute of Technology in 1977, the M.S. degree from Madras University in 1979, and the Ph.D. degree from Concordia University in 1990.

He has 20 years of experience in the networking industry, where he worked for Cisco, Nortel, Alcatel, and Fujitsu and was responsible for advanced technology products from concept to delivery. For the last five years, he has been researching on the prevention of denial-of-service and spam attacks in voice over Internet Protocol (VoIP) networks. He is currently an Associate Professor with the Department of Computer Science and Engineering, University of North Texas (UNT), Denton. He is the Founding Director of the Network Security Laboratory at UNT, the objective of which is to study the problems and issues related to next-generation networks. He is also the Director of the Center for Information and Computer Security, UNT. Prior to UNT, he was a Technology Director with Netrake, where he was the Architect of the redundancy mechanism for VoIP firewalls. He has cochaired three workshops on VoIP security. His additional experience includes being a Technical Director in IPMobile (acquired by Cisco), where he was instrumental in the wireless/IP product concept, architecture, design, and delivery. His research focus is on detecting spam, network security, and next-generation networks. In addition to more than 100 research papers, he has authored several Requests For Comments related to Multiprotocol Label Switching, SS7 over IP, and routing. Due to his innovative work, Cisco and Alcatel were granted a total of 20 patents; another 15 are pending.


**João W. Cangussu** (M'98) received the B.S. degree in computer science from the Federal University of Mato Grosso do Sul, Campo Grande, Brazil, in 1990, the M.S. degree in computer science from the University of São Paulo, São Carlos, Brazil, in 1993, and the Ph.D. degree in computer science from Purdue University, West Lafayette, IN, in 2002.

After that, he was with the Department of Computer Science, University of Texas at Dallas, Richardson, until 2010 when he has been with the Search Team at Microsoft Corporation, Redmond, WA. His research interests include software testing, software process modeling and control, adaptive systems, and the application of control theory of computer-science-related problems.