Behavioral Entropy of a Cellular Phone User

Santi Phithakkitnukoon[1], Husain Husna[2], and Ram Dantu[3]

[1]santi@unt.edu, Department of Comp. Sci. & Eng., University of North Texas
[2]hjh0036@unt.edu, Department of Comp. Sci. & Eng., University of North Texas
[3]rdantu@unt.edu, Department of Comp. Sci. & Eng., University of North Texas

Abstract The increase of advanced service offered by cellular networks draws lots of interest from researchers to study the networks and phone user behavior. With the evolution of Voice over IP, cellular phone usage is expected to increase exponentially. In this paper, we analyze the behavior of cellular phone users and identify behavior signatures based on their calling patterns. We quantify and infer the relationship of a person's randomness levels using information entropy based on the location of the user, time of the call, inter-connected time, and duration of the call. We use real-life call logs of 94 mobile phone users collected at MIT by the Reality Mining Project group for a period of nine months. We are able to capture the user's calling behavior on various parameters and interesting relationship between randomness levels in individual's life and calling pattern using correlation coefficients and factor analysis. This study extends our understanding of cellular phone user behavior and characterizes cellular phone users in forms of randomness level.

1 Introduction

Mobile phone has moved beyond being a mere technological object and has become an integral part of many people's social lives. This has had profound implications on both how people as individuals perceive communication as well as in the patterns of communication of humans as a society. In this paper we try to capture the behavior of phone users based on their calling patterns and infer trend of behavior dependencies using techniques such as Entropy, principal factor analysis and correlation function. We present a new method for precise measurement of randomness of phone user based on their calling patterns such as location of the call, talk time, calling time and interconnected time and infer relationship among them.

Recently there has been increasingly growing interests in the field of mobile social networks analysis, but due to the unavailability of data, there have been far fewer studies. The Reality Mining Project at Massachusetts Institute of Technology (MIT) [1] has made publicly available large datasets from their projects. We implement our techniques on the Reality Mining dataset which was collected over nine months by monitoring the cell phone usage of 94 participants. The information collected in the call logs includes user IDs (unique number representing a mobile phone user), time of call, call direction (incoming and outgoing), incoming call description (missed,

accepted), talk time, and tower IDs (location of phone users). These 94 phone users are students, professors, and staffs.

Using purely objective data first time the researchers can get an accurate glimpse into human behaviors. Our interest in this data set is to study the behavior of the phone user using information theory, data mining, and data reduction techniques.

In [2], the authors attempted to quantify the amount of predictable structure in an individual's life using entropic metric and discovered that people who live high-entropy lives tend to be more random or less predictable than people who live low-entropy lives. This raises the question about how this entropy-based randomness level is related to the randomness level in calling behavior. Does it mean that people who have high-entropy lives also have high-entropy calling patterns? To answer this question, we find it interesting to study the relationship between the randomness level in individuals life and calling pattern.

The main contribution of this paper is to infer the relationship between the randomness levels in behavior of the phone users in a cellular network. We believe that this work can also be extended to predict what services that are suitable for the user.

The rest of this paper is structured as follows: Section 2 carries out the randomness level computation based on entropy. Section 3 discusses the randomness computation result and the relationship among them. The paper is concluded in section 4 with a summary and an outlook on future work.

2 Randomness Level Computation

While individual phone user's calling behavior is random, some users might be more predictable than others. Being more predictable can also mean being less random. To quantify the randomness or amount of predictable structure in an individual calling pattern, the information entropy can be used.

The information entropy or Shannon's entropy is a measure of uncertainty of a random variable. The information entropy as given in Eq. (1) was introduced by Shannon [3], where X is a discrete random variable, $x \in X$, and the probability mass function $p(x) = Pr\{X = x\}$.

$$H(X) = -\sum_{x} p(x) \log_2 p(x). \tag{1}$$

The calling pattern can be observed from the calling time, inter-connected time (elapsed time between two adjacent call activities), and talk time (duration of call). Let C, I, and T be random variables representing calling time, inter-connected time, and talk time respectively. The entropy of calling time can be calculated by Eq. (2).

$$H(C) = -\sum_{c=1}^{24} p(c) \log_2 p(c), \tag{2}$$

Table 1 Result of Correlation Coefficient

	$H(L)$	$H(C)$	$H(I)$	$H(T)$
$H(L)$	1.0000	0.4651	-0.4695	-0.4642
$H(C)$	0.4651	1.0000	-0.2218	-0.3502
$H(I)$	-0.4695	-0.2218	1.0000	0.2197
$H(T)$	-0.4642	-0.3502	0.2197	1.0000

where the probability $p(c)$ is a ratio of the number of calls during c^{th} hour slot to the total number of calls of all time slots (N).

Similarly, the entropy of inter-connected time can be calculated by Eq. (3) where $p(i)$ is a ratio of the number of inter-connected time whose value is in the interval $[i-1,i)$ to $N-1$.

$$H(I) = -\sum_{i} p(i)\log_2 p(i). \tag{3}$$

Likewise, the entropy of the talk time is given by Eq. (4) where $p(t)$ is a ratio of the talk time whose value is in the interval $[t-1,t)$ to N.

$$H(T) = -\sum_{t} p(t)\log_2 p(t). \tag{4}$$

By the same token, the randomness in the individual life's schedule (location), $H(L)$ can also be quantified using information entropy which is defined in Eq. (1).

3 Result and Analysis

Based on our real-life call logs of 94 users, we infer the relationship between the randomness based on the underlying parameters by computing the correlation coefficient [4]. Correlation coefficient is a number between -1 and 1 which measures the degree to which two random variables are linearly related. A correlation coefficient of 1 implies that there is perfect linear relationship between the two random variables. A correlation coefficient of -1 implies that there is inversely proportional relationship between the two random variables. A correlation coefficient of zero implies that there is no linear relationship between the variables. As a preliminary result shown in Table 1, it can be observed that the randomness based on location($H(L)$) and calling time($H(C)$) show high correlation as well as the $H(I)$ and $H(T)$ pair.

Next, we perform factor analysis in order to further study the relationship of the randomness levels (entropy) based on the underlying parameters. The main application of factor analysis is: (1) to reduce the number of variables and (2) to detect structure in the relationship between variables, that is to classify variables [5]. In our analysis we use it for both the purposes. The flow diagram of the principal factor analysis is shown in Fig. 1.

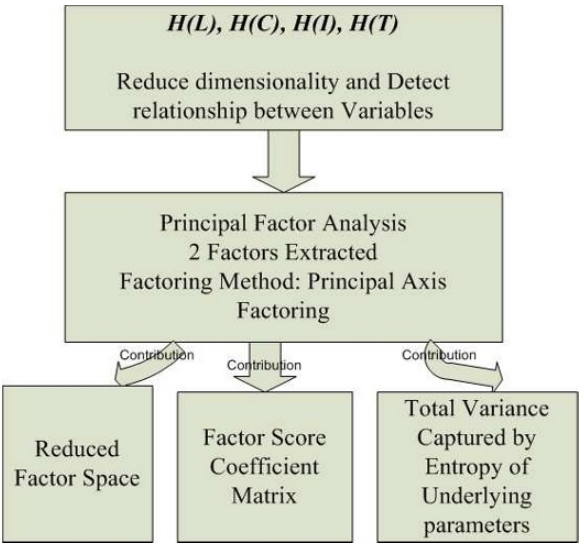

Fig. 1 Flow diagram for principal factor analysis on calculated entropy.

Table 2 Total Variance Explained

Factor	Initial Eigen Values			Extraction		
	Total	Variance(%)	Cumulative(%)	Total	Variance(%)	Cumulative(%)
1	1.59	39.95	39.95	0.92	23.20	23.20
2	1.02	25.61	65.57	0.29	7.26	30.46
3	0.73	18.24	83.81	-	-	-
4	0.64	16.18	100.00	-	-	-

Two principal factors are selected based on the Scree plot [7]. The principal factor plot of the entropy based on four parameters lying on the first and second factor is shown in Fig. 1. It can be observed that the $H(L)$ and $H(C)$ are positively lying on the first factor whereas the $H(I)$ and $H(T)$ are positively lying on the second factor. Since the first and second factor are orthogonal i.e., uncorrelated, one can notice two established relations; one is between $H(L)$ and $H(C)$, and the other one is between $H(I)$ and $H(T)$.

Factor analysis generally is used to encompass both principal components and principal factor analysis. The Eigen values for a given factor measures the variance in all the variables which is accounted for by that factor as stated in Table 2. If a factor has a low eigen value, then it is contributing little to the explanation of variances in the variables and may be ignored as redundant with more important factors.

Eigen value is not the percent of variance explained but rather a measure of amount of variance in relation to total variance (since variables are standardized to have means of 0 and 1, total variance is equal to the number of variables).

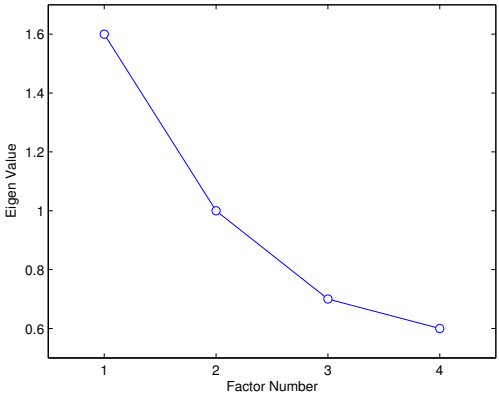

Fig. 2 Scree plot.

Initial eigen values and eigen values after extraction (extracted sums of squared loadings) are same for Principal Component Analysis (PCA) extraction [6], but for factor analysis eigen values after extraction will be lower than their initial counterparts.

Scree plot was developed by Cattell [7] for selecting the number of factors to be retained in order to account for most of the variation. In our analysis, based on Kaiser's criterion [8] the first two factors whose eigen values are greater than 1 (as listed in Table 3) are selected based on the scree plot shown in Fig. 2.

The plot of the entropy based on four parameters lying on the first and second factor is shown in Fig. 3. It can be observed that the entropy based on location and calling time are positively lying on the first factor whereas the entropy based on inter-connected time and talk time are positively lying on the second factor. Since the first and second factor are orthogonal i.e., uncorrelated, one can notice two established relations; (1) between entropy based on location and calling time and (2) entropy based on inter-connected time and talk time.

The scatter plots in Fig. 4 also confirm our findings by showing the proportional relationships between pairs $(H(L), H(C))$ and $(H(I), H(T))$, and inversely proportional relationships among other pairs. The trend (linear-fitting) line is shown in red to emphasize the direction of the relationship, directly proportional (increasing) or inversely proportional (decreasing). Note that the linear fitting is obtained by the least square fitting method [9].

The results based on the correlation coefficients, factor analysis, and scatter plots tell us that there is a high correlation in the randomness in phone user's location and calling time, as well as high correlation in the randomness in phone user's interconnected time and talk time. This draws the conclusion of our study that phone users who have higher randomness in mobility tend to be more variable in time of making calls but less variable in time spent talking on the phone and the time between connection (idle time). By the same token, the phone users who spend

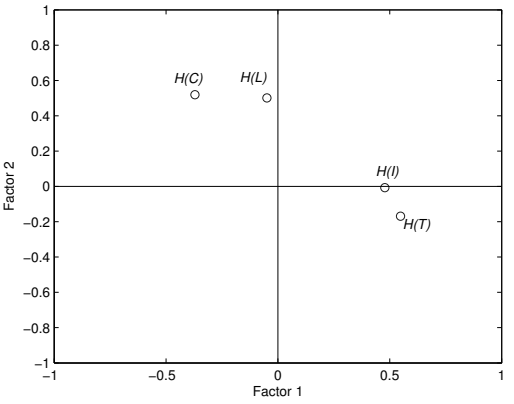

Fig. 3 Principal factor plot.

higher random amount of time talking on the phone (connected time) tend to also be more variable in idle time but not less random in mobility and time of initiating the calls.

We believe that this finding can also be useful for the phone service providers in offering right plans for the right customers based on customer's calling behavior, e.g. suppose that a customer has increasingly high randomness in mobility, service provider might offer this customer a whenever-minute plan which would fit his calling pattern (high $H(L)$ implies high $H(C)$).

4 Conclusion

In this paper, we have presented and analyzed cellular phone user behavior in forms of randomness level using information entropy based on user's location, time of call, inter-connected time, and duration of call. We are able to capture the relationship of the user's randomness level based on the underlying parameters by utilizing the correlation coefficient and factor analysis.

Based on our study, the user's randomness level based on location has high correlation to the randomness level in time of making phone calls and vice versa. Our study also shows that the randomness level based on user's inter-connected time has a high correlation to the randomness level in time spent talking on each phone call.

A knowledge of the randomness levels of a phone user behavior and their relationships extends our understanding in the pattern of user behavior. We believe that this work can also be extended to predict what services that are suitable for the user. This study will also be useful for the future research in this area. As our future direction, this study will be applied to quantify the presence information in terms of willingness level of a phone user in accepting a call.

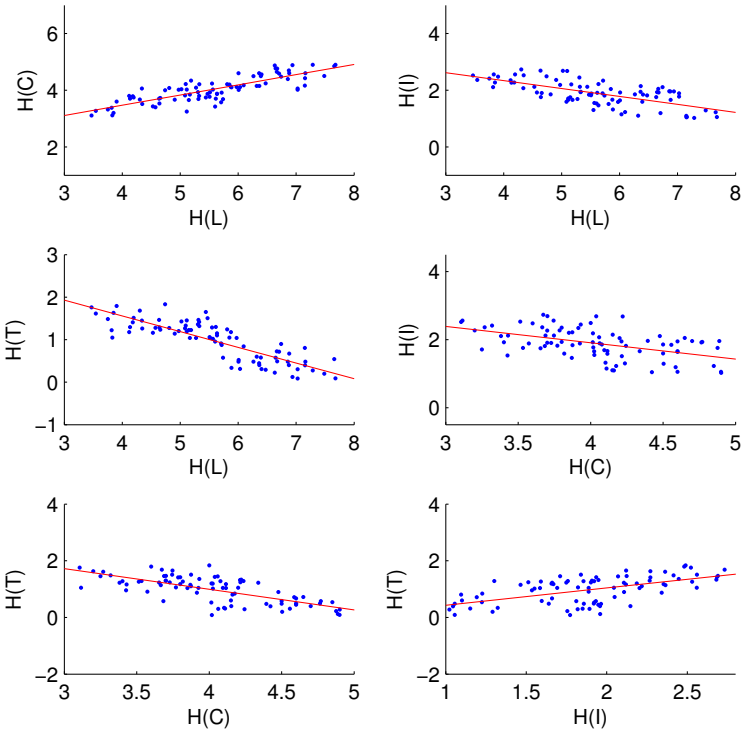

Fig. 4 Scatter plots showing relationships among $H(L), H(C), H(I), and H(T)$ with the linear trend lines.

Acknowledgments.

This work is supported by the National Science Foundation under grants CNS-0627754, CNS-0619871, and CNS-0551694. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

We would like to thank the Reality Mining Project group, particularly Dr. Nathan Eagle and Dr. Alex (Sandy) Pentland of MIT Media Lab for providing us the valuable datasets.

References

1. Massachusetts Institute of Technology: Reality Mining Project. Available: http://reality.media.mit.edu/
2. Eagle, N., and Pentlend, A.: Reality Mining: Sensing Complex Social Systems. Personal and Ubiquitous Computing, Vol. 10, No. 4, 2006.
3. Shanon, C.E.: A mathematical theory of communication. Bell System Technical Journal, Vol. 27, pp.379-423 and 623-656, July and October 1948.
4. Cohen, J.: Statistical power analysis for the behavioral sciences. 2nd ed. Lawrence Erlbaum Associates, Hillsdale, NJ, 1988.
5. Jolliffe, I. T.: Principal Component Analysis. 2nd ed. Springer Science+Business Media, 1986, New York USA.
6. Eagle, N., Pentland, A., and Lazer, D.: Infering social network structure using mobile phone data. Proc. of National Academy of Sciences, 2006.
7. Cattell, R. B., and Vogelmann, S.: A Comprehensive trial of the scree and KG criteria for determining the number of factors. Mult. Behav. Res., Vol. 12, pp. 289-325, 1977.
8. Kaiser, H.F.: The application of electronic computers to factor analysis. Educ. Psychol. Meas. 20, 141e151, 1960.
9. De Groen, P.: An introduction to total least squares. in Nieuw Archief voor Wiskunde, Vierde serie, deel 14, 1996, pp. 237-253.