Physica A 393 (2014) 513–518

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Analysis of the origin of predictability in human communications

Lin Zhang [a], Yani Liu [a], Ye Wu [a,b,*], Jinghua Xiao [a,b]

[a] School of Science, Beijing University of Posts and Telecommunications, Beijing 100876, China

[b] State Key Lab of Information Photonics and Optical Communications, Beijing University of Posts and Telecommunications, Beijing 100876, China

HIGHLIGHTS

- Human communication partner is predictable in e-mails, short-messages and cell-phone calls.
- People with fewer communication partners have higher predictability.
- The intrinsic pattern and the burst are the origin of predictability in human communications.

ARTICLE INFO

Article history:
Received 29 May 2013
Received in revised form 15 August 2013
Available online 30 August 2013

Keywords:
Human dynamics
Communication patterns
Predictability
Entropy

ABSTRACT

Human behaviors in daily life can be traced by their communications via electronic devices. E-mails, short messages and cell-phone calls can be used to investigate the predictability of communication partners' patterns, because these three are the most representative and common behaviors in daily communications. In this paper, we show that all the three manners have apparent predictability in partners' patterns, and moreover, the short message users' sequences have the highest predictability among the three. We also reveal that people with fewer communication partners have higher predictability. Finally, we investigate the origin of predictability, which comes from two aspects: one is the intrinsic pattern in the partners sequence, that is, people have the preference of communicating with a fixed partner after another fixed one. The other aspect is the burst, which is communicating with the same partner several times in a row. The high burst in short message communication pattern is one of the main reasons for its high predictability, the intrinsic pattern in e-mail partners sequence is the main reason for its predictability, and the predictability of cell-phone call partners sequence comes from both aspects.

1. Introduction

Human dynamics is a complex and important topic, because human behaviors are usually considered to be random and unpredictable. The massive amount of data that social media generates has promoted the study of human communication behavior. One of the hot topics in human dynamics is the study of inter-event time distribution. Barabási et al. found that the inter-event time distribution is power-law in human communications [1]. The scaling properties and kinds of explanations for this heavy-tailed power laws are also investigated by others [2–4]. Power-law inter-event time distribution indicates that human dynamics has memory effect compared with the traditional light-tailed exponential distribution. This important discovery opens up a new perspective and more and more researchers are interested in human dynamics because of the power-law feature in theories and potential applications such as human dynamics in epidemics [5–10], and so on.

* Corresponding author at: School of Science, Beijing University of Posts and Telecommunications, Beijing 100876, China. Tel.: +86 18911311020.
E-mail address: wuye@bupt.edu.cn (Y. Wu).

0378-4371/$ – see front matter
http://dx.doi.org/10.1016/j.physa.2013.08.054

Table 1
Detailed format of short messages data.

User's ID	Partner's ID	Sent time
A	$A1$	2012/02/01,12:05:05
A	$A1$	2012/02/01,12:10:05
A	$A2$	2012/02/01,12:13:05
...	...	...
B	$B1$	2012/02/01,12:01:05
...	...	...

Another important topic in human dynamics recently is the predictability of human dynamics. Finding out the predictive social behaviors could be used for better understanding of human's behavior. Song et al. found that the human mobility patterns are 93% predictable based on the whereabouts of people recorded by mobile phones [11]. Domenico et al. studied the interdependence and predictability of human movement and social ties to improve the accuracy of movement forecasting [12]. Qin et al. found that the distribution of typical places that persons visited is heavy-tailed power law using the extracted smart-phone based data [13]. The regularities and randomness of mobility are also analyzed. Moreover, the predictability of communication partners is also a hot subtopic in this area. Inspired by the method of measuring predictability in Song [11], Takaguchi et al. investigated the predictability of conversation partners in face-to-face interactions within an organization [14]. They found that the uncertainty about the next partner in the sequence of one's conversation partners given the current partner is decreased by 28.4% on average. For online social interactions, Wang et al. showed that individual interactions are more predictable when users acted on their own rather than attended group activities [15].

In this paper, three databases generated from e-mails, short messages and cell-phone calls are used to study the predictability of human communication interactions in daily life. We wonder whether human communications in daily life can be predicted and what are the differences in the predictabilities of different communication manners. Compared with the face-to-face data used in [14] and the online data used in [15], these three databases are more representative and more common in people's daily communication behaviors. We found that the uncorrelated entropy of the user's sequence is larger than the conditional entropy in all the three cases. This is an obvious evidence that the communication partners are not totally random and can be predicted to some extent no matter which device people choose for communication. By comparison, the short message database has the highest predictability in partners pattern of the three. We have also found that people with fewer communication partners are more predictable. Finally, the origin of predictability is revealed. Reasons for predictability comes from two aspects. One is the intrinsic pattern in the partners sequence, that is, people have the preference of communicating with a fixed partner after another fixed one. The other aspect is the burst, which is communicating with the same partner several times in a row. The high burst in short message communication pattern is the main reason for its high predictability, the intrinsic pattern in e-mail partners sequence is the main reason for its predictability, and the predictability of cell-phone call partners sequence comes from both the two aspects. These results can be helpful for both government and companies. The results should be helpful for mobile service companies. It is also helpful in our daily life such as helping the police catching thieves by measuring the thieves' predictability based on communication histories.

This paper is organized as follows: In Section 2, the detailed information of our data are introduced. The definition of different entropies, statistical results for patterns, and reasons for predictability are presented in Section 3. Finally, our conclusions are given in Section 4.

2. Description of original data

Three sets of human communication data are analyzed in our paper. Data set D_1 is the emails that are sent by $N_1 = 714$ individuals in 82 days at a university in Europe [16]. Data sets D_2 and D_3 are the short messages and cell-phone calls over a month period with 380 thousand users. The format of the data we use is shown in Table 1 (take the short message data set D_2 for illustration). The data includes three columns. The first column is the ID of each user, say User A for the first user, which is an artificial one and is provided for illustration purposes. The second column is the ID of User A's communication partners and the last column is the time accurate to seconds that the message was sent from User A to the partners. All the ID information has been transformed for privacy protection. In order to compare the three databases together, 1000 users in D_2 and D_3 are selected randomly from the total users. The results for larger samples have been verified to have the same statistical properties with the 1000-user sample. The information that we focus on is the sequence of each user's partner in time order, neglecting the sent time.

3. Methods and results

In order to calculate the predictability of the communication partner sequence, information theory is used inspired by Refs. [11,14,15]. Define three different entropies to measure each user's communication patterns:

$$H_A^0 = \log_2 N_A,$$