Predictability of conversation partners

Taro Takaguchi, Mitsuhiro Nakamura, Nobuo Sato, Kazuo Yano, and Naoki Masuda

preprint arXiv:1104.5344
社会ネットワーク中での会話パターンの予測可能性

“Predictability of conversation partners”
preprint arXiv:1104.5344

高口太朗, 中村光宏, 佐藤信夫, 矢野和男, 増田直紀
Conventional assumptions

Modeling human behavior such as

- Mobility patterns in the physical space by random (or Lévy) walk
- Temporal order of selecting interaction partners by Poisson process

Actual behavior: to what degree random / deterministic?
Predictability of mobility patterns (Song et al., 2010)

Mobility patterns are largely predictable.

Method: measuring the entropy of the sequence of locations of each cell-phone user

\[ T = \{ X_1, \cdots, X_t, X_{t+1}, \cdots, X_L \} \]

How about other components of human behavior?
Predictability of interaction partners

interacts with

"partner sequence"
\{2, 3, 2, 2, 3, 1\}

- Characterize individuals in a social organization
- Related to epidemics and information spreading
Data set: face-to-face interaction log in an office

Recording from two offices in Japanese companies
- subjects: 163 individuals
- period: 73 days
- total conversation events: 51,879 times

We used the data collected by World Signal Center, Hitachi, Ltd., Japan.

Name tag with an infrared module
(The Business Microscope system)

http://www.hitachi-hitec.com/jyouhou/business-microscope/
Details of data set

- A conversation event = communication between close tags
- Conversation events: undirected
- Time resolution = one minute
  - Multiple events initiated within the same minute
  → determine their order at random

<table>
<thead>
<tr>
<th>time stamp</th>
<th>ID1</th>
<th>ID2</th>
<th>duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>2009-01-01 14:13</td>
<td>1</td>
<td>3</td>
<td>1 min</td>
</tr>
<tr>
<td>2009-01-01 14:15</td>
<td>1</td>
<td>6</td>
<td>4 min</td>
</tr>
<tr>
<td>2009-01-01 14:15</td>
<td>5</td>
<td>26</td>
<td>1 min</td>
</tr>
<tr>
<td>2009-01-01 14:18</td>
<td>2</td>
<td>5</td>
<td>1 min</td>
</tr>
<tr>
<td>2009-01-01 14:19</td>
<td>1</td>
<td>13</td>
<td>4 min</td>
</tr>
<tr>
<td>2009-01-01 14:19</td>
<td>3</td>
<td>4</td>
<td>1 min</td>
</tr>
<tr>
<td>2009-01-01 14:24</td>
<td>1</td>
<td>22</td>
<td>1 min</td>
</tr>
<tr>
<td>2009-01-01 14:26</td>
<td>1</td>
<td>22</td>
<td>7 min</td>
</tr>
<tr>
<td>2009-01-01 14:26</td>
<td>3</td>
<td>6</td>
<td>1 min</td>
</tr>
</tbody>
</table>

(c) \{3, 6, 13, 22, 22, ...\}
Three entropy measures (cf. Song et al., 2010)

- for individual $i$ who has $k_i$ partners in the recording period,

Random entropy: interaction in a totally random manner

$$H_i^0 = \log_2 k_i$$

Uncorrelated entropy: heterogeneity among $P_i(j)$

$$H_i^1 = - \sum_{j \in N_i} P_i(j) \log_2 P_i(j)$$

$P_i(j)$: the probability to talk with $j$

Conditional entropy: second-order correlation

$$H_i^2 = - \sum_{j \in N_i} P_i(j) \sum_{\ell \in N_i} P_i(\ell|j) \log_2 P_i(\ell|j)$$

$P_i(\ell|j)$: to talk with $\ell$ immediately after with $j$
cf. Entropy

“The uncertainty of the random variable”

\[ H(P) \equiv - \sum_{\omega \in \Omega} P(\omega) \log_2 P(\omega) \]

Ex.) coin toss

\( \omega \in \{\text{head}, \text{tail}\} \quad \text{\( P(\text{head}) = p \)} \)

\[ H(p) = -p \log_2 p - (1 - p) \log_2 (1 - p) \]

\( p = 0 : \text{always tail} \)
Example 1: random

Suppose that

\[ P_i(1) = \frac{1}{2}, \quad P_i(2) = \frac{1}{3}, \quad P_i(3) = \frac{1}{6} \]

\[ \begin{align*}
    P_i(1|\ast) &= P_i(1), \\
    P_i(2|\ast) &= P_i(2), \\
    P_i(3|\ast) &= P_i(3).
\end{align*} \]

\[
H_i^0 = \log_2 3
\]

\[
H_i^1 = - \sum_{j=1}^{3} P_i(j) \log_2 P_i(j) = \frac{2}{3} + \frac{1}{2} \log_2 3
\]

\[ H_i^2 = H_i^1 < H_i^0 \]
Example 2: predictable

Suppose that
\[ P_i(1) = \frac{1}{2}, \quad P_i(2) = \frac{1}{3}, \quad P_i(3) = \frac{1}{6} \]

\[ P_i(\ell|j) = \begin{bmatrix} 0 & 1 & 1 \\ 2/3 & 0 & 0 \\ 1/3 & 0 & 0 \end{bmatrix} \]

\[ H_i^0 = \log_2 3 \]

\[ H_i^1 = \frac{2}{3} + \log_2 3 \]

(same as example 1)

\[ H_i^2 = \left( \frac{1}{2} \log_2 3 - \frac{1}{3} \right) < H_i^1 < H_i^0 \]
Quantifying the predictability

Mutual information of the partner sequence

\[ I_i = H_i^1 - H_i^2 \]

Interpretation:
The information about the NEXT partner that is earned by knowing the PREVIOUS partner

\[ \begin{array}{c}
0 \\
I_i \\
H_i^1 \\
\end{array} \]

random | predictable
Aim of the study

- Examine the predictability of conversation partners
  - similar to that of mobility patterns
- “Predictability” = the mutual information of the partner sequence
- Data set: face-to-face interaction log from two offices in Japan

\[ I_i = H_i^1 - H_i^2 \]
$I_i = H_{i}^{1} - H_{i}^{2} \gg 0$, regardless of the value of $H_{i}^{1}$

→ Partner sequences are predictable to some extent.
Significance of $I_i$

Null hypothesis: 
“$I_i$ is positive only because of the small data size.”

Compare $I_i$ with the value obtained from shuffled partner sequences

![Graph showing empirical $I_i$ against shuffled sequences](image)
Main cause of the predictability

The bursty human activity pattern: characterized by the long-tailed inter-event intervals

“\(i\) tends to talk with \(j\) within a short period after talking with \(j\).”

\[P_{>}(\tau)\] of a typical individual
Purpose: examine the contribution of the bursty activity

Partner sequence in a given day

extract the sequences with each partner

with 1

with 2

with 3
Next, shuffle the intervals within each partner

<table>
<thead>
<tr>
<th>Partner</th>
<th>Intervals</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>( \tau_{1}^{(2)} ), ( \tau_{1}^{(4)} ), ( \tau_{1}^{(3)} ), ( \tau_{1}^{(1)} )</td>
</tr>
<tr>
<td>2</td>
<td>( \tau_{2}^{(3)} ), ( \tau_{2}^{(2)} ), ( \tau_{2}^{(1)} )</td>
</tr>
<tr>
<td>3</td>
<td>( \tau_{3}^{(3)} ), ( \tau_{3}^{(1)} ), ( \tau_{3}^{(2)} )</td>
</tr>
</tbody>
</table>

merge

Calculate \( I_{i}^{\text{burst}} \) of this shuffled sequence
Result of Test 1

Generate 100 shuffled sequence

Contribution of burstiness \( \approx 80\% \)
Test 2

**Purpose:** examine the predictability after omitting the burstiness

$$\{2, 3, 2, 3, 1\}$$

Calculate $I_i^{\text{merge}}$ of the merged sequence
Result of Test 2

Shuffle the merged sequence with replacement conditioned that no partner ID appears successively

\[ I^\text{merge}_i \]

is significantly large.

\[ \rightarrow \] Predictability remains without the effect of the burstiness.

Error bars: 99% confidential intervals of shuffled sequences

Original \( I^\text{merge}_i \)
Summary so far

- Partner sequence: predictable to some extent
- Main cause of predictability: the bursty activity pattern
  - Predictability remains after omitting the burstiness.
Predictability $I_i$ depends on individuals

Histogram of $I_i$

Depends on $i$’s position in the social network?
Conversation network

Undirected and weighted network

Node : individual
Link : a pair of individuals having at least one event
Link weight : The total number of events for the pair

Node attributions

Degree $k_i$
Strength $s_i = \sum_j w_{ij}$
Mean weight $\overline{w}_i = s_i / k_i$
Correlation between $I_i$ and node attributions

- No correlation (Left)
- $R = -1.698 \times 10^{-3}$

- Negative (Center)
- $R = -0.5111$

- Negative (Right)
- $R = -0.6533$
Hypotheses

- Fix $k_i$, small $w_i$ $\rightarrow$ abundance of weak links (with small weights)

- Hypothesis about links
  - “Strength of weak ties” (Granovetter, 1973)

- Hypothesis about nodes

```
$w \geq 5$
```

```
$w \leq 4$
```
Weak links offer non-redundant contacts.
Weak links tend to connect different communities.

“bridge”
Hypotheses

- Fix $k_i$, small $w_i \rightarrow$ abundance of weak links (with small weights)

☐ Hypothesis about links
  - Weak links connect different communities.

☐ Hypothesis about nodes
  - Individuals connecting different communities $\rightarrow$ large $I_i$
  - Individuals concealed in a community $\rightarrow$ small $I_i$
Hypothesis about links

**Purpose:** quantify the degree of being concealed in a community

Relative neighborhood overlap of a link (Onnela *et al.*, 2007)

\[ O_{ij} = \frac{|\text{Intersection of the neighbors of } i \text{ and } j|}{|\text{Union of the neighbors of } i \text{ and } j| - 2} \]

\[ O_{ij} = 0 \]

\[ O_{ij} = 1/2 \]

\[ O_{ij} = 1 \]
Hypothesis about links

“Strength of weak ties” hypothesis; a link with large $w_{ij}$ tends to have a large $O_{ij}$

$\rightarrow O_{ij}$ increases with $w_{ij}$

$\langle O \rangle_w : O_{ij}$ averaged over the links with $w_{ij} \leq w$

Consistent with the hypothesis
Hypotheses

- Fix $k_i$, small $w_i \rightarrow$ abundance of weak links (with small weights)

✓ Hypothesis about links
  - Weak links connects different communities.

☐ Hypothesis about nodes
  • Individuals connecting different communities $\rightarrow$ large $I_i$
  • Individuals concealed in a community $\rightarrow$ small

$w \geq 5$
$w \leq 4$
Hypothesis about nodes

Purpose: quantify the degree of being concealed in a community

Clustering coefficient of a node (Watts and Strogatz, 1998)

\[ C_i = \frac{\# \text{ triangles including } i}{k_i(k_i - 1)/2} \]

\[ C_i = 0 \quad C_i = 1/2 \quad C_i = 1 \]
Abundance of triangles

- 3-clique percolation community (Palla *et al.*, 2005)

- Hierarchical structure (Ravász and Barabási, 2003)
  Nodes connecting communities → small $C_i$
Clustering coefficient with link-weight threshold

$C_i(w_{thr})$: after eliminating the links with $w_{ij} < w_{thr}$

Original CN

Expectation: triangles persist around individuals with small $I_i$

$C_i(w_{thr})$ for a fixed $w_{thr}$
Hypothesis about nodes

■: Pearson correlation coefficient between $I_i$ and $C_i(w_{\text{thr}})$

○: Partial correlation coefficient between $I_i$ and $C_i(w_{\text{thr}})$ with $k_i(w_{\text{thr}})$ and $s_i(w_{\text{thr}})$ fixed
Hypotheses

- Fix $\kappa_i$, small $\overline{w}_i \rightarrow$ abundance of weak links (with small weights)

✓ Hypothesis about links

  - Weak links connects different communities.

✓ Hypothesis about nodes

  • Individuals connecting different communities $\rightarrow$ large $I_i$
  • Individuals concealed in a community $\rightarrow$ small $I_i$
Conclusions

- **Predictability of partner sequence**
  - Face-to-face interaction log in offices in Japanese companies

- **Conversation partners: predictable to some extent**
  - A main cause: the bursty activity pattern
  - Significant predictability after omitting the burstiness

- **Dependence on the individual’s position in the conversation network**
  - “Strength of weak ties”
  - INTER-community individuals $\rightarrow$ large $I_i$
  - INTRA-community individuals $\rightarrow$ small $I_i$

Preprint available electronically

Taro Takaguchi, Mitsuhiro Nakamura, Nobuo Sato, Kazuo Yano, and Naoki Masuda,
“Predictability of conversation partners”,
preprint arXiv:1104.5344