# 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

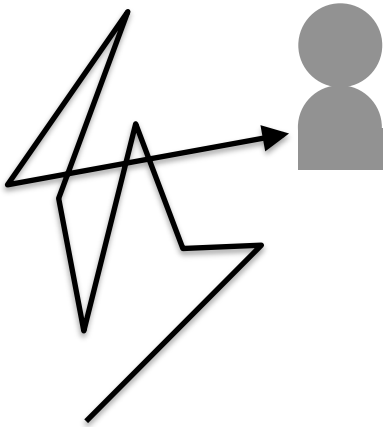
“Predictability of conversation partners”
preprint arXiv:1104.5344

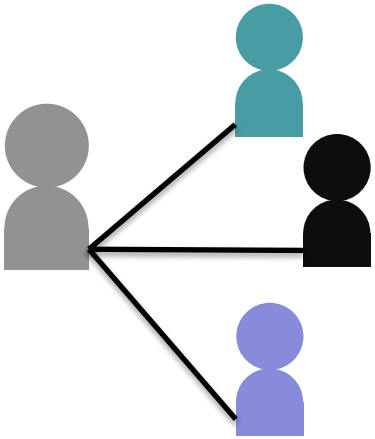
高口太朗, 中村光宏, 佐藤信夫, 矢野和男,
増田直紀

# Conventional assumptions

Modeling human behavior such as

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





Actual behavior: to what degree random / deterministic ?

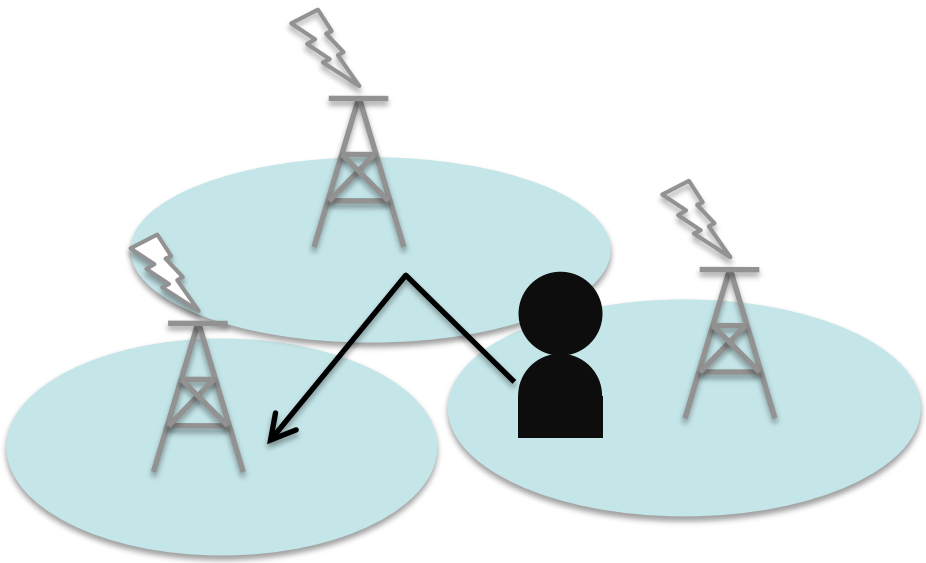
# Predictability of mobility patterns (Song *et al*., 2010)

Mobility patterns are largely predictable.

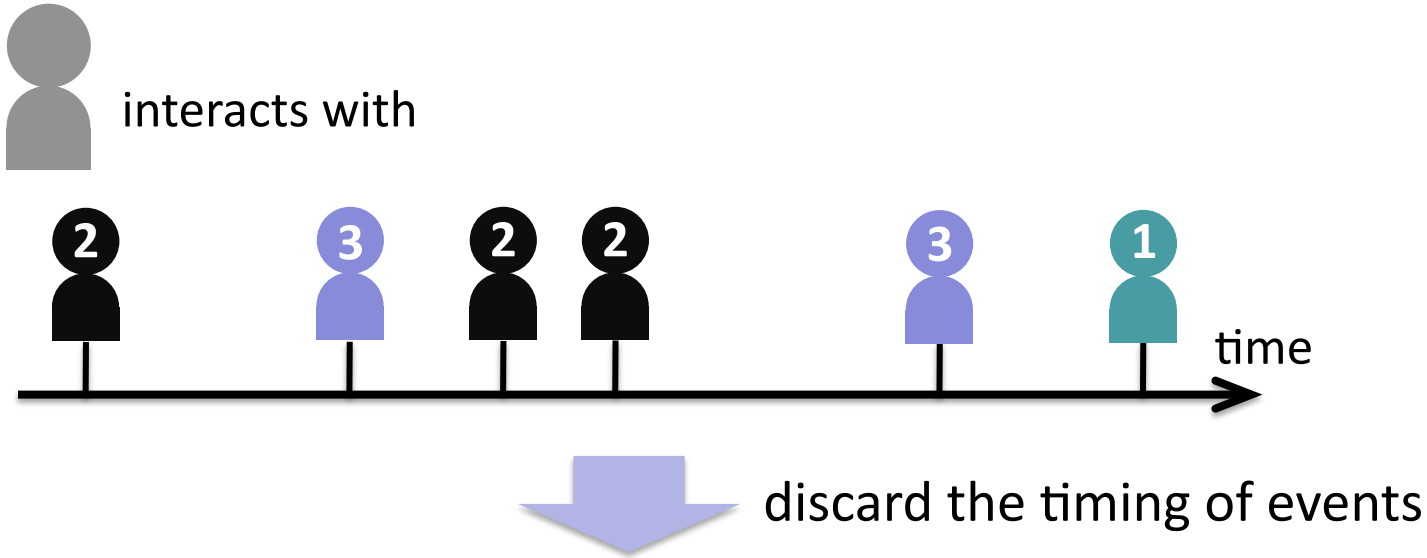
Method: measuring the entropy of
<u>the sequence of locations of each cell-phone user</u>

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

How about other components of human behavior?



# Predictability of interaction partners




"partner sequence"

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

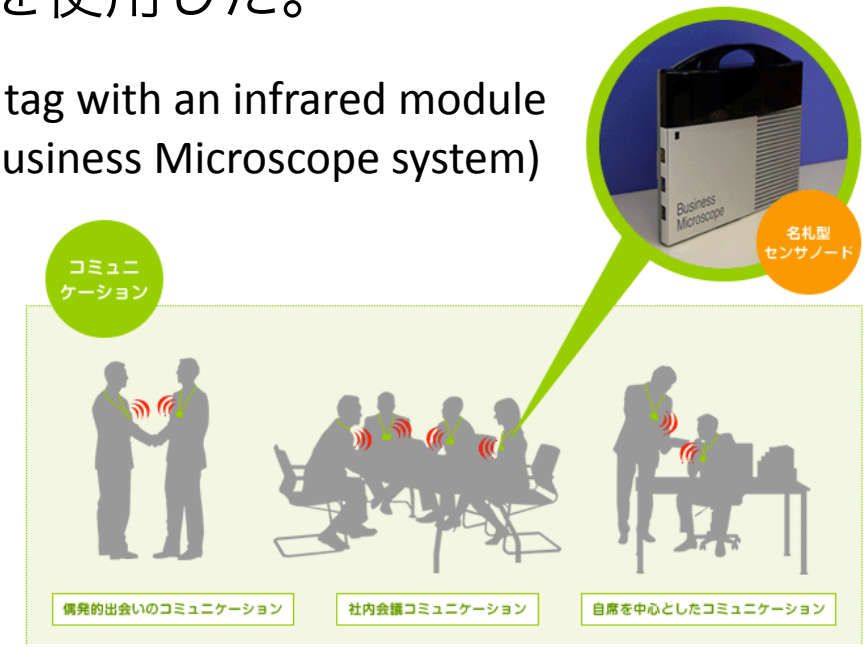
- Characterize individuals in a social organization
- Related to epidemics and information spreading

# 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.
(株)日立製作所 ワールドシグナルセンタにより収集された
データセットを使用した。

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

(a)

| time stamp | ID1 | ID2 | duration |
|---|---|---|---|
| 2009-01-01 14:13 | 1 | 3 | 1 min |
| 2009-01-01 14:15 | 1 | 6 | 4 min |
| 2009-01-01 14:15 | 5 | 26 | 1 min |
| 2009-01-01 14:18 | 2 | 5 | 1 min |
| 2009-01-01 14:19 | 1 | 13 | 4 min |
| 2009-01-01 14:19 | 3 | 4 | 1 min |
| 2009-01-01 14:24 | 1 | 22 | 1 min |
| 2009-01-01 14:26 | 1 | 22 | 7 min |
| 2009-01-01 14:26 | 3 | 6 | 1 min |
| ⋮ | | | |

→

(b)

| time stamp | ID1 | ID2 | duration |
|---|---|---|---|
| 2009-01-01 14:13 | 1 | 3 | 1 min |
| 2009-01-01 14:15 | 1 | 6 | 4 min |
| 2009-01-01 14:19 | 1 | 13 | 4 min |
| 2009-01-01 14:24 | 1 | 22 | 1 min |
| 2009-01-01 14:26 | 1 | 22 | 7 min |
| ⋮ | | | |

→

(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 $\mathrm{P}_i(j)$

$$H_i^1 = -\sum_{j \in \mathcal{N}_i} \mathrm{P}_i(j) \log_2 \mathrm{P}_i(j)$$

$\mathrm{P}_i(j)$: the probability to talk with $j$

Conditional entropy: second-order correlation

$$H_i^2 = -\sum_{j \in \mathcal{N}_i} \mathrm{P}_i(j) \sum_{\ell \in \mathcal{N}_i} \mathrm{P}_i(\ell|j) \log_2 \mathrm{P}_i(\ell|j)$$

$\mathrm{P}_i(\ell|j)$: to talk with $\ell$ immediately after with $j$

# cf. Entropy

"The uncertainty of the random variable"

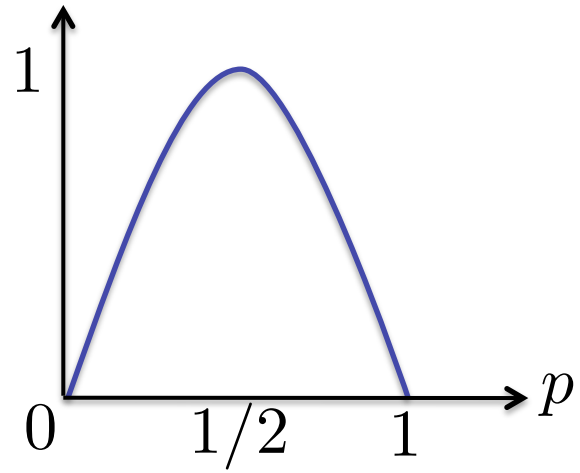
$$H(P) \equiv -\sum_{\omega \in \Omega} P(\omega) \log_2 P(\omega)$$

Ex.) coin toss

$$\omega \in \{\text{head},\ \text{tail}\}$$

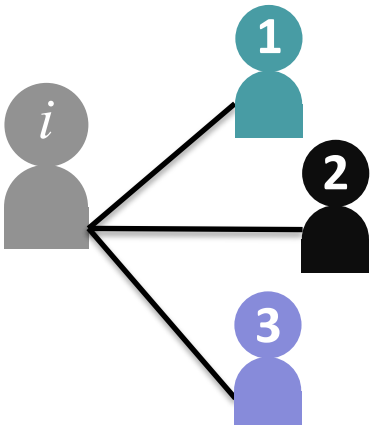
$$P(\text{head}) = p$$

$$H(p) = -p \log_2 p - (1-p) \log_2 (1-p)$$




$p = 0$ : always tail

# Example 1: random




Suppose that

$$\mathrm{P}_i(1) = 1/2,\ \mathrm{P}_i(2) = 1/3,\ \mathrm{P}_i(3) = 1/6$$

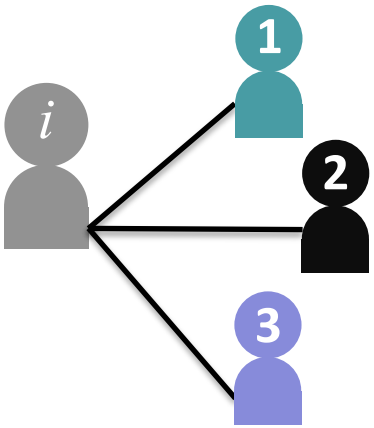
$$\begin{cases} \mathrm{P}_i(1|*) = \mathrm{P}_i(1), \\ \mathrm{P}_i(2|*) = \mathrm{P}_i(2), \\ \mathrm{P}_i(3|*) = \mathrm{P}_i(3). \end{cases}$$

$$H_i^0 = \log_2 3$$

$$H_i^1 = -\sum_{j=1}^{3} \mathrm{P}_i(j) \log_2 \mathrm{P}_i(j) = \frac{2}{3} + \frac{1}{2} \log_2 3$$

$$\boxed{H_i^2 = H_i^1 < H_i^0}$$

# Example 2: predictable




Suppose that

$$\mathrm{P}_i(1) = 1/2,\ \mathrm{P}_i(2) = 1/3,\ \mathrm{P}_i(3) = 1/6$$

$$\mathrm{P}_i(\ell|j) \qquad j$$

$$\ell \begin{bmatrix} 0 & 1 & 1 \\ 2/3 & 0 & 0 \\ 1/3 & 0 & 0 \end{bmatrix}$$

$$H_i^0 = \log_2 3$$

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

(same as example 1)

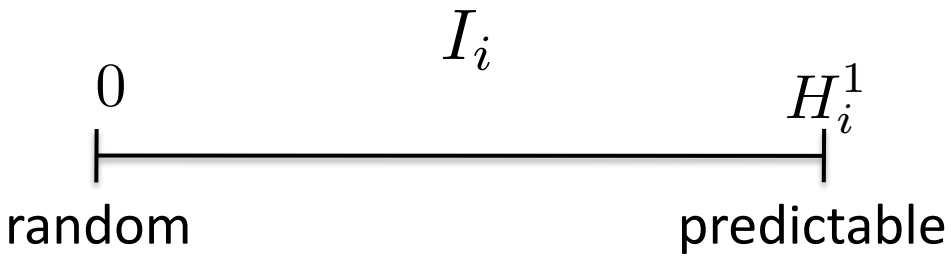
$$\boxed{H_i^2 = \left(\frac{1}{2}\log_2 3 - \frac{1}{3}\right) < H_i^1 < H_i^0}$$

# Quantifying the predictability

Mutual information of the partner sequence

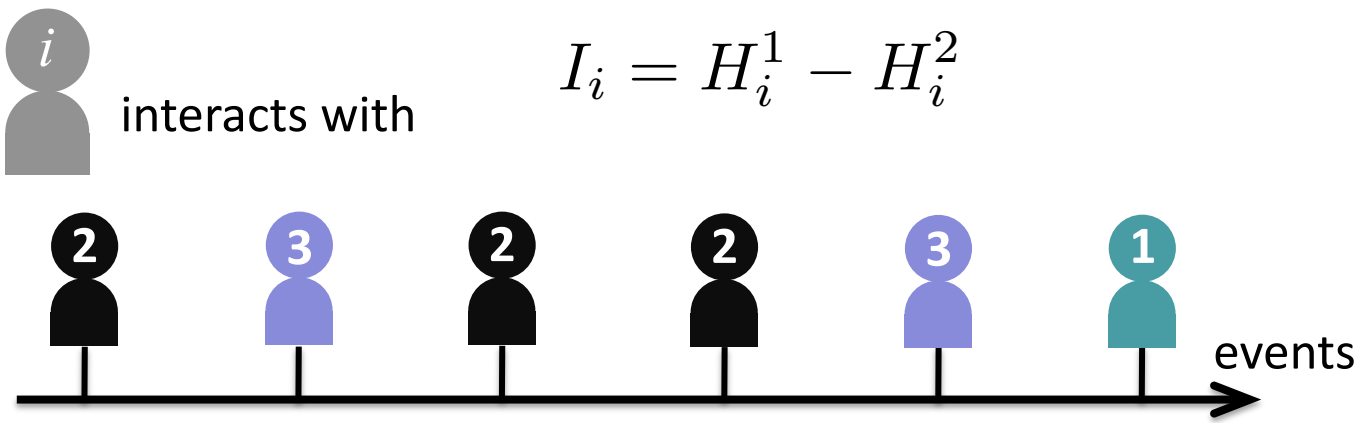
$$I_i = H_i^1 - H_i^2$$

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



# 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

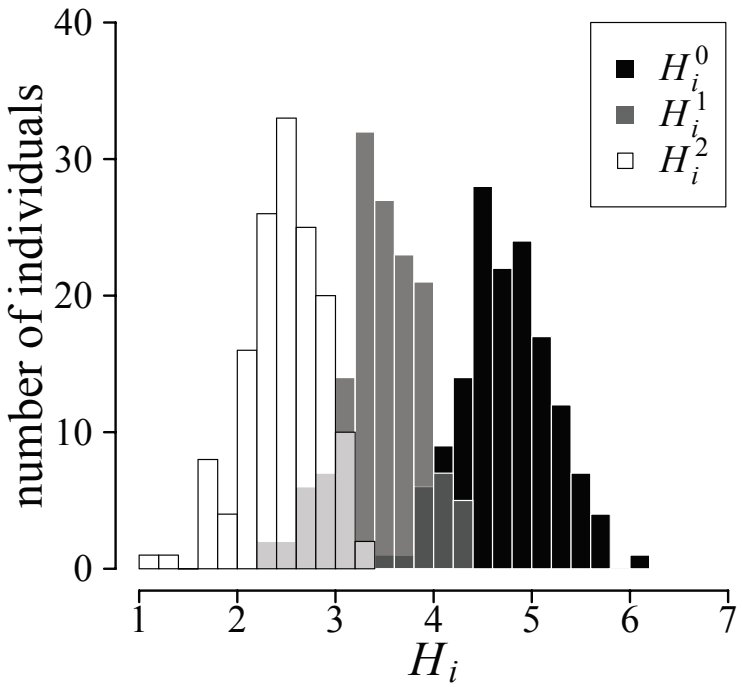


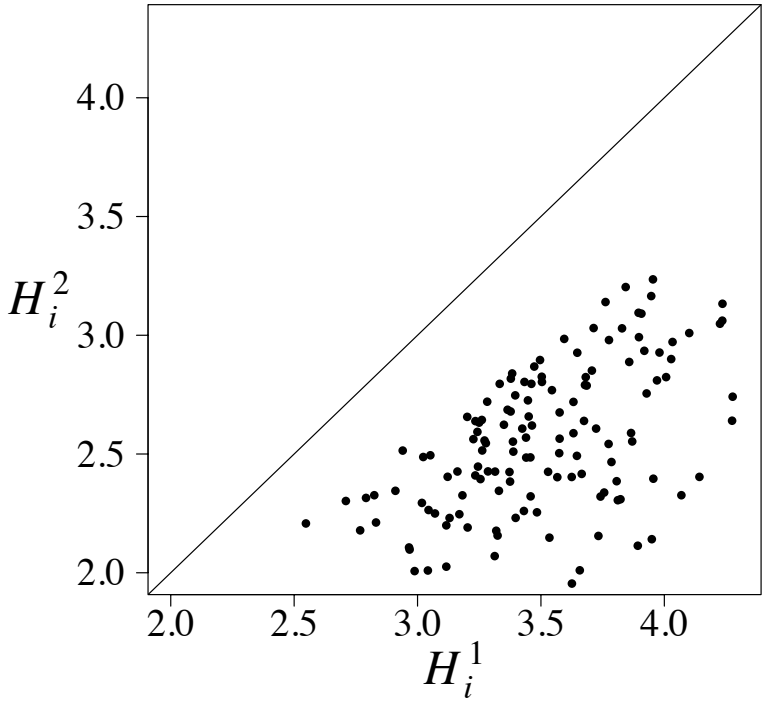

$$I_i = H_i^1 - H_i^2$$

# Histogram of the entropies

$I_i = H_i^1 - H_i^2 \gg 0$, regardless of the value of $H_i^1$

→ Partner sequences are predictable to some extent.

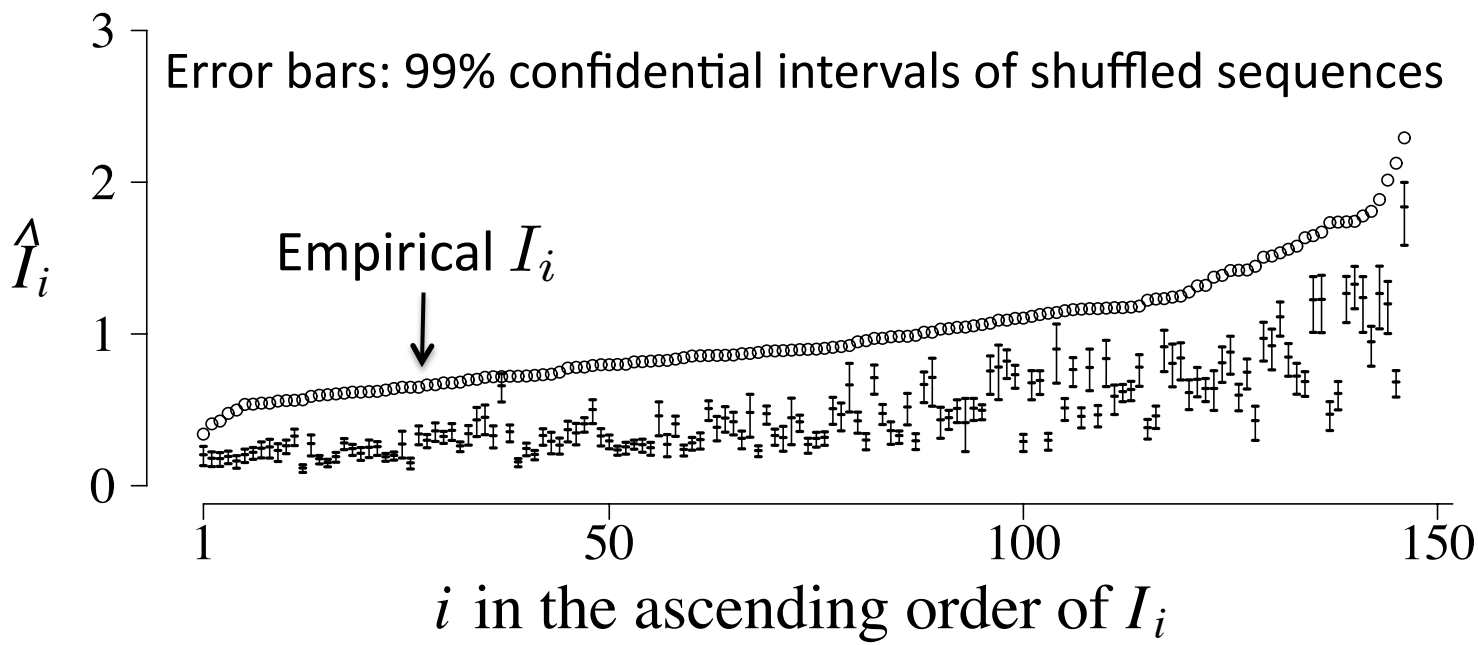


# Significance of $I_i$

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

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

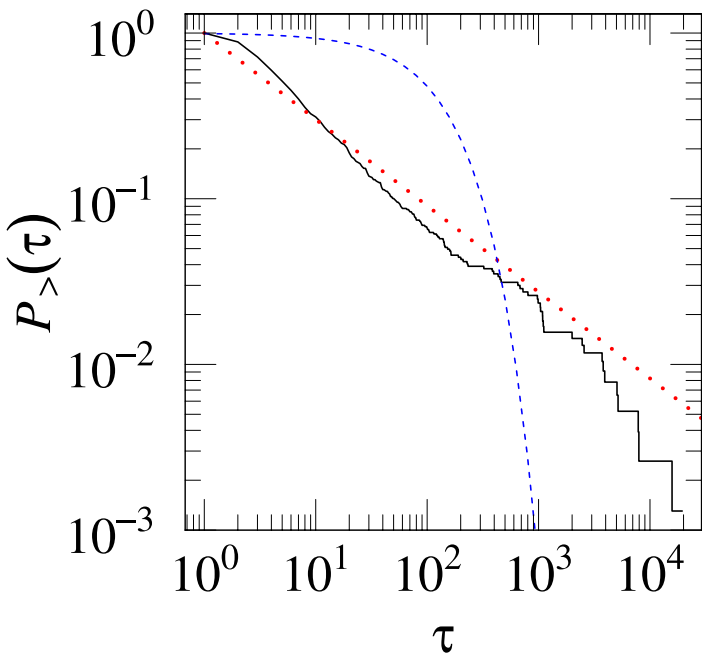


# Main cause of the predictability

<u>The bursty human activity pattern</u>:
characterized by the long-tailed inter-event intervals

"$i$ tends to talk with $j$ within a short period after talking with $j$."

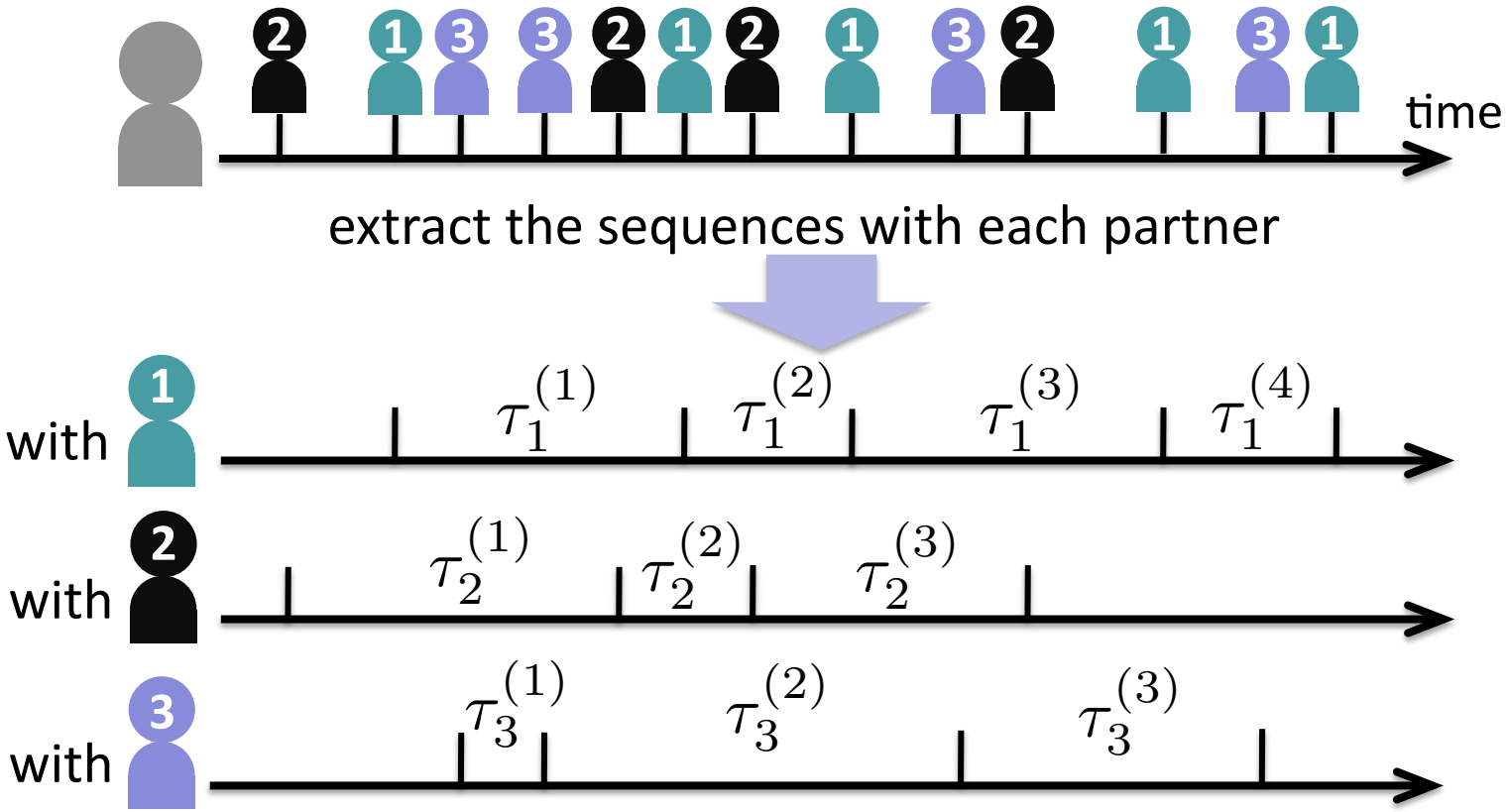

$P_>(\tau)$ of a typical individual



# Test 1 (1)

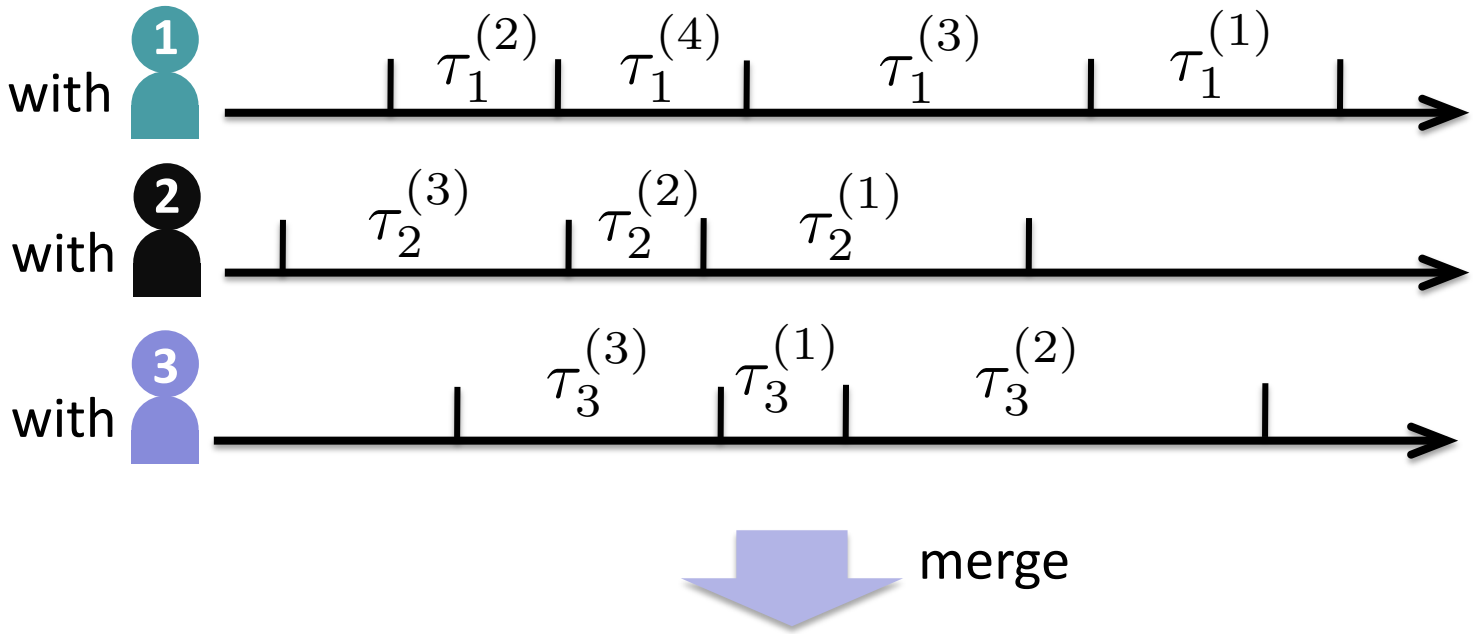
Purpose: examine the contribution of the bursty activity

Partner sequence in a given day

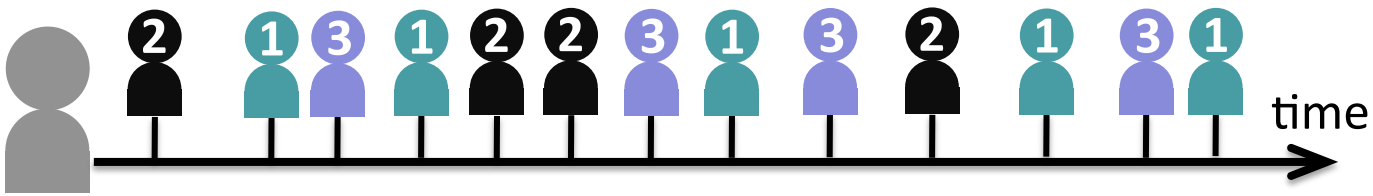


# Test 1 (2)

Next, shuffle the intervals within each partner



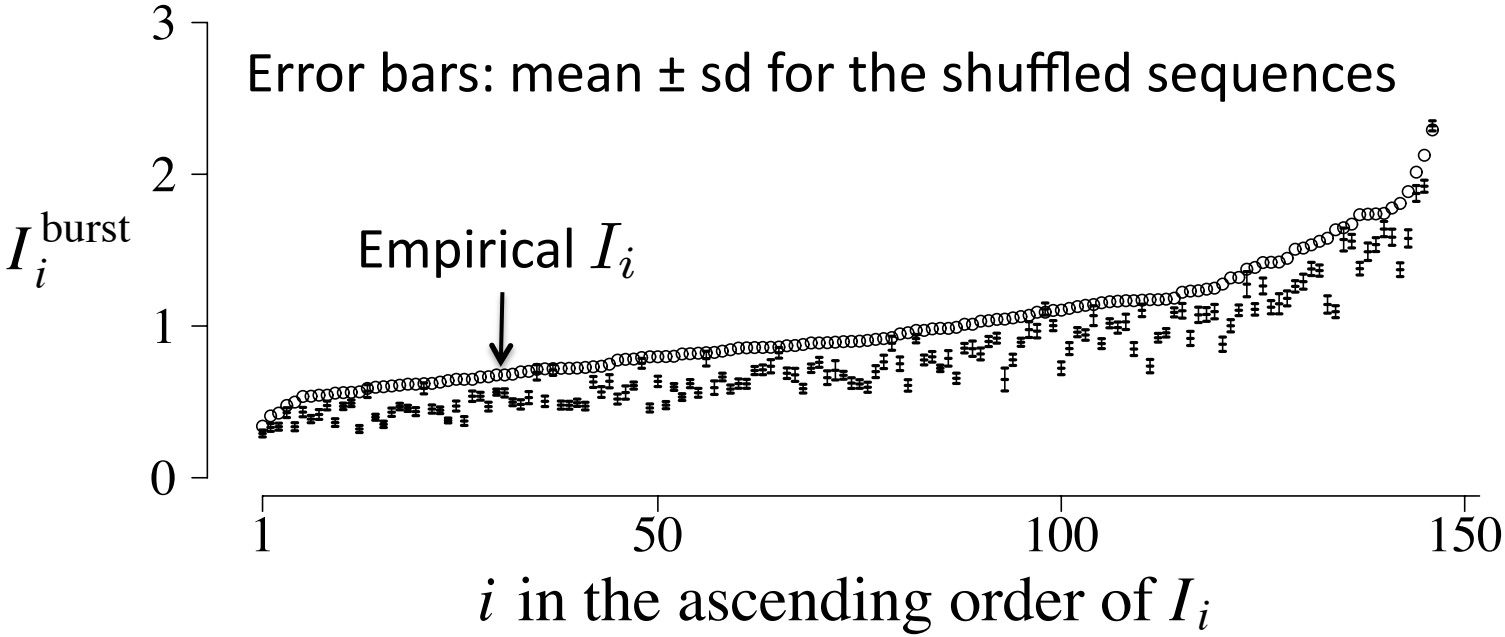

Calculate $I_i^{\mathrm{burst}}$ of this shuffled sequence



# Result of Test 1

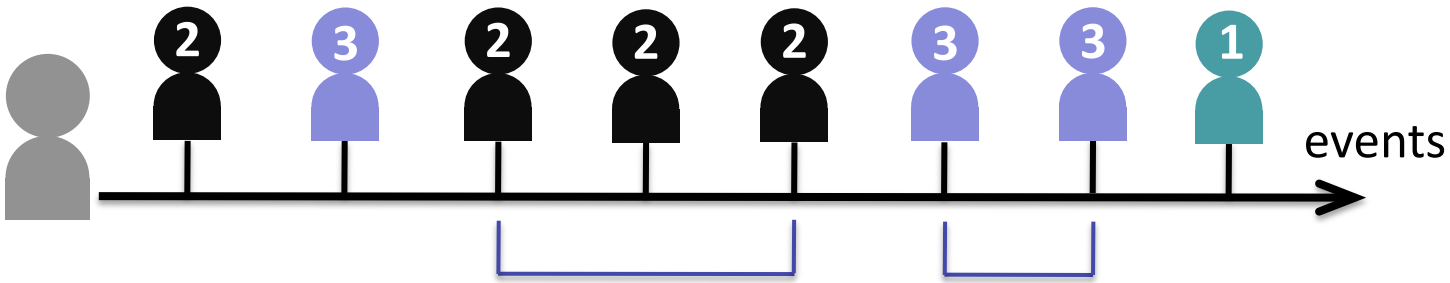
Generate 100 shuffled sequence

Contribution of burstiness $\approx 80\%$



# Test 2

<u>Purpose</u>: examine the predictability after omitting the burstiness




merge into one

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

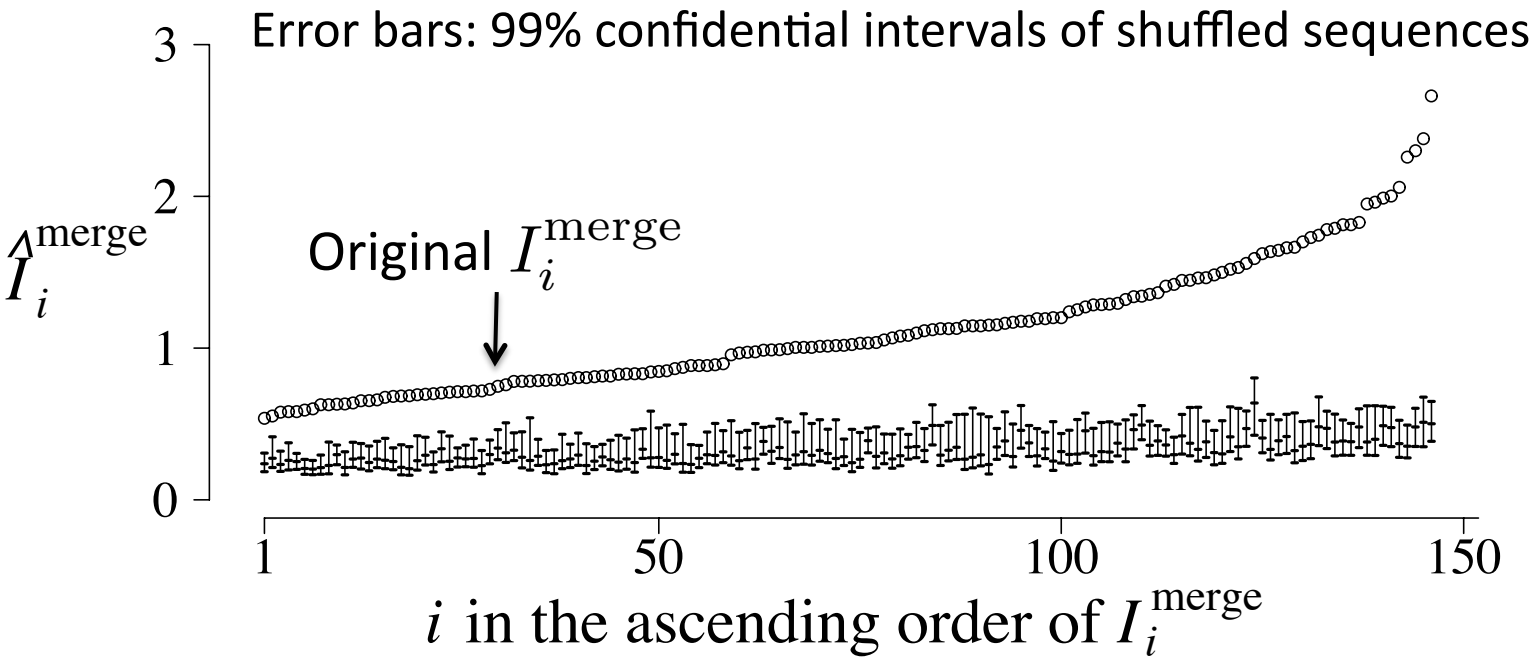
Calculate $I_i^{\mathrm{merge}}$ of the merged sequence

# Result of Test 2

Shuffle the merged sequence with replacement
<u>conditioned that no partner ID appears successively</u>

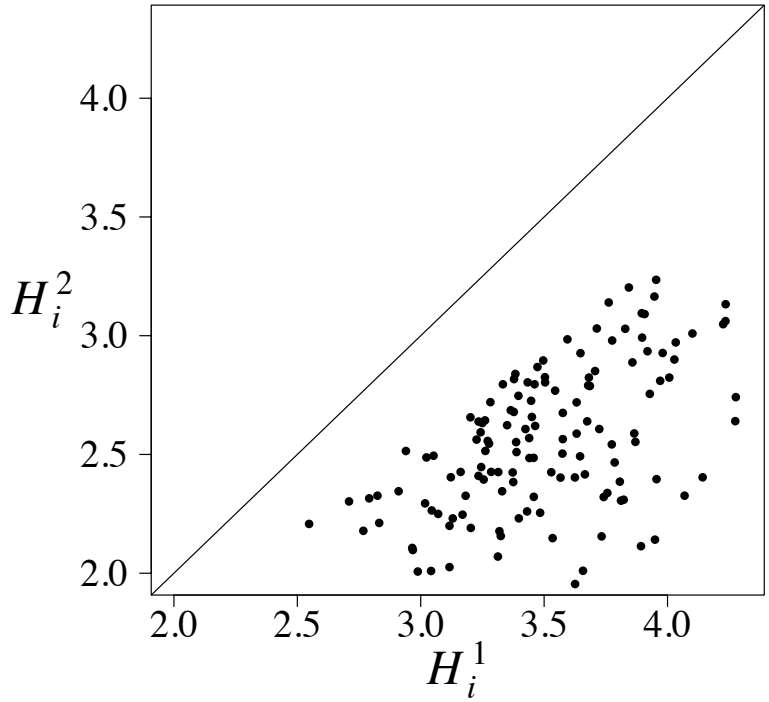
$I_i^{\mathrm{merge}}$ is significantly large.

→ Predictability remains without the effect of the burstiness.



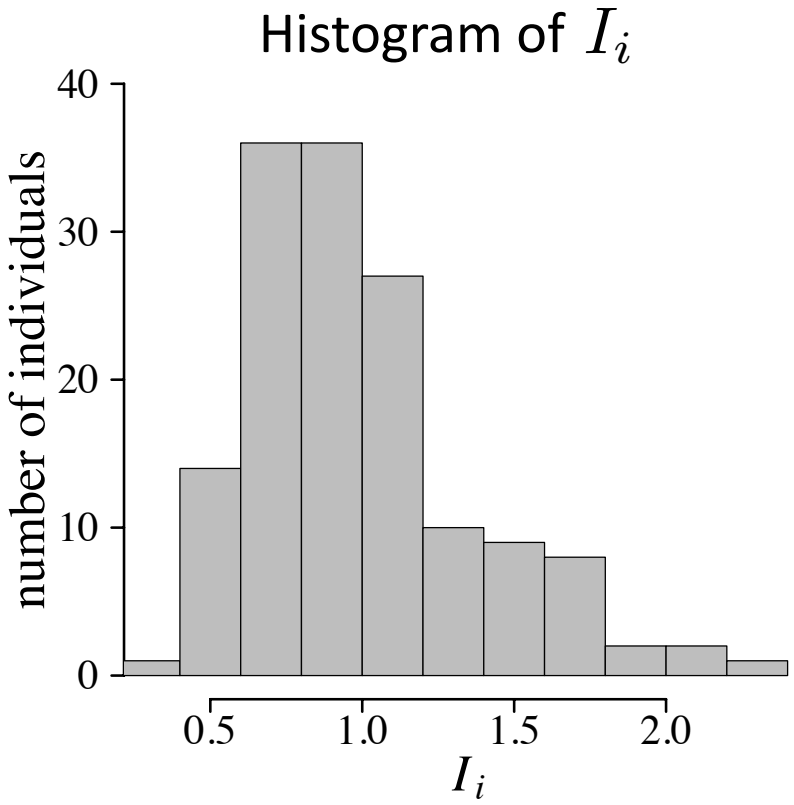
# Summary so far

- 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

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

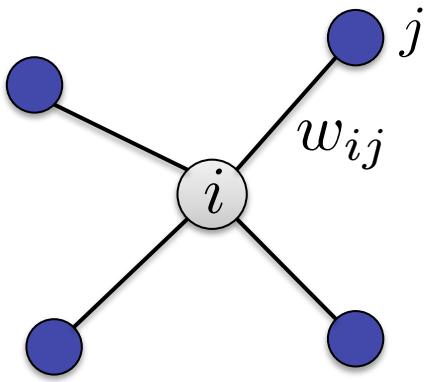
Link weight : The total number of events for the pair

<u>Node attributions</u>

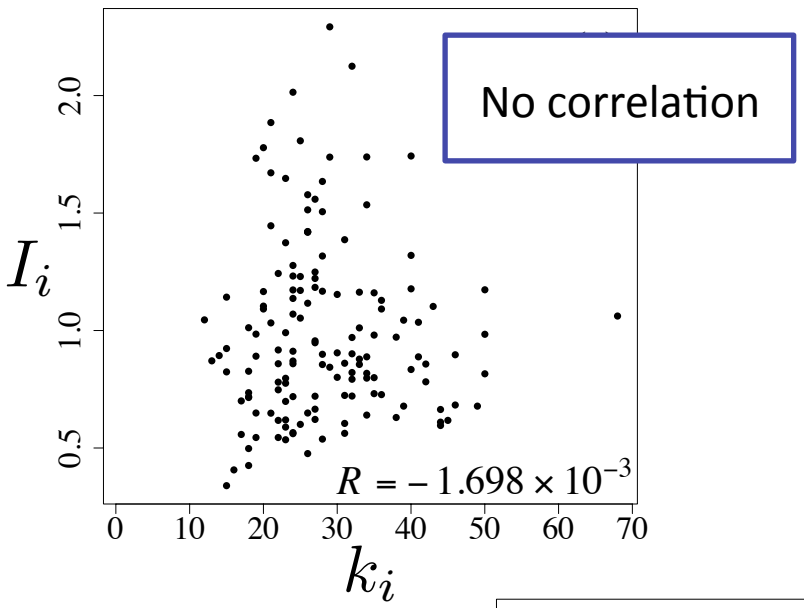
Degree $k_i$

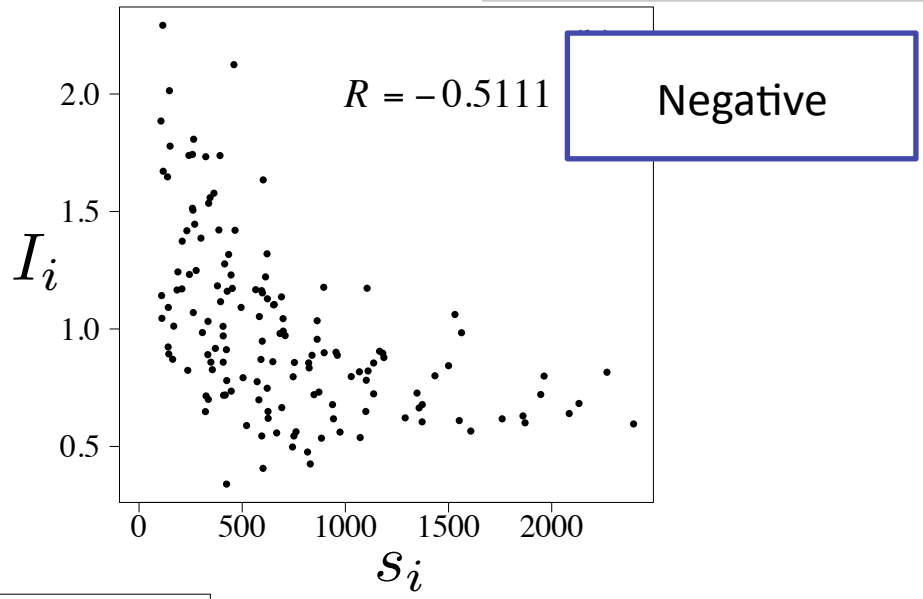
Strength $s_i = \sum_j w_{ij}$

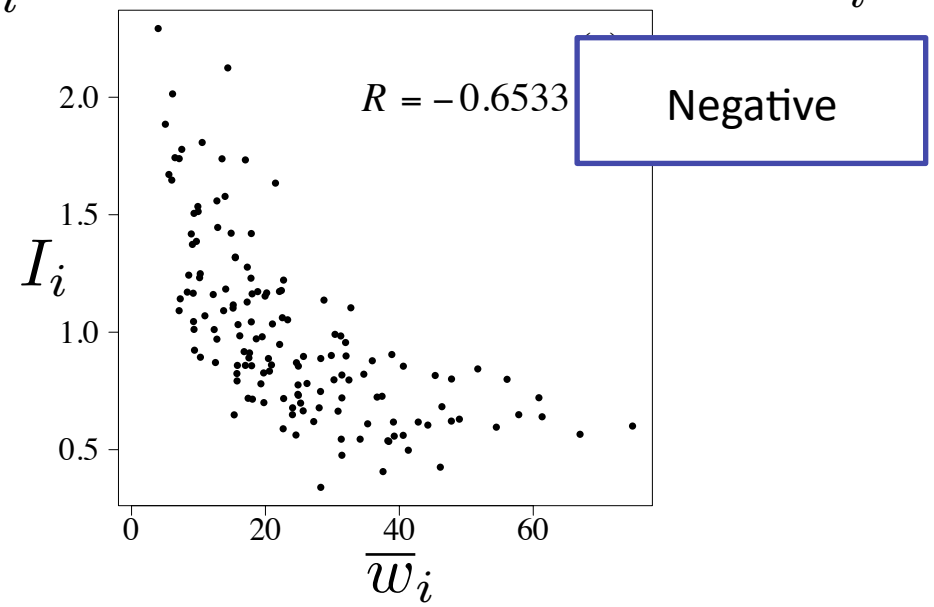
Mean weight $\overline{w}_i = s_i / k_i$



# Correlation between $I_i$ and node attributions



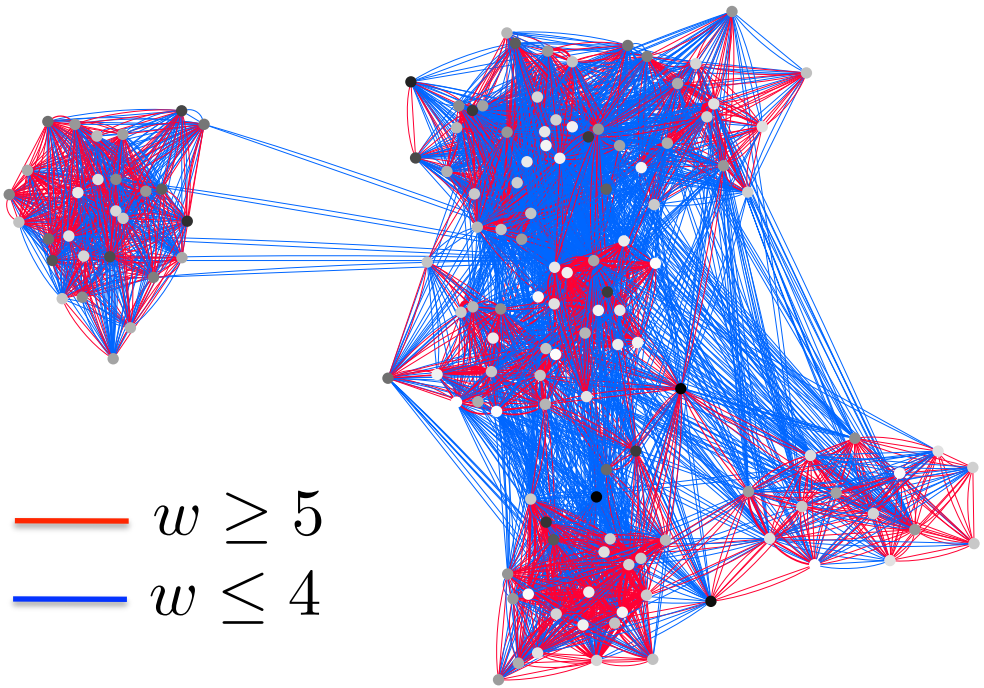
# Hypotheses

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

☐ <u>Hypothesis about links</u>

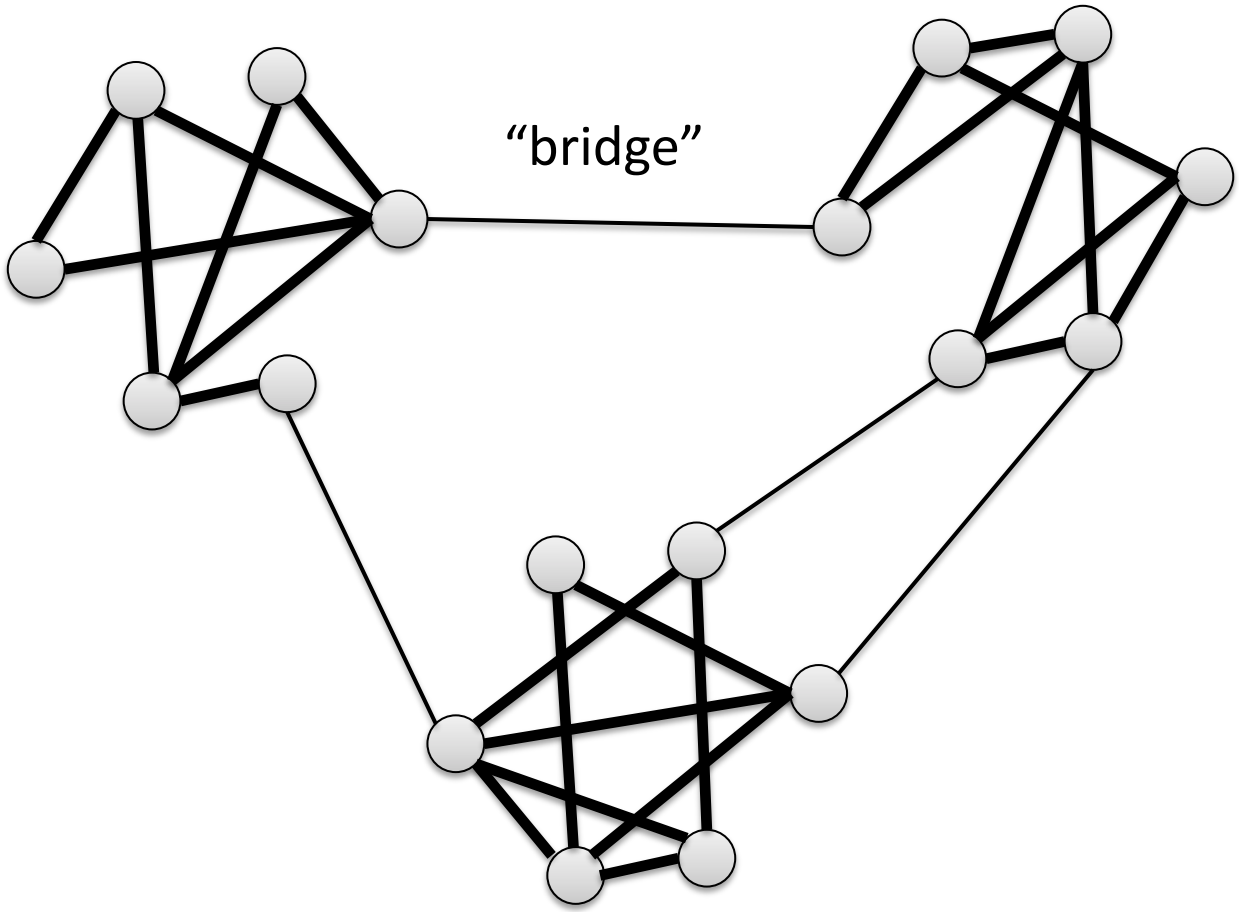
- "Strength of weak ties" (Granovetter, 1973)

☐ <u>Hypothesis about nodes</u>



# Strength of weak ties (Granovetter, 1973)

Weak links offer non-redundant contacts.
Weak links tend to connect different communities.



# Hypotheses

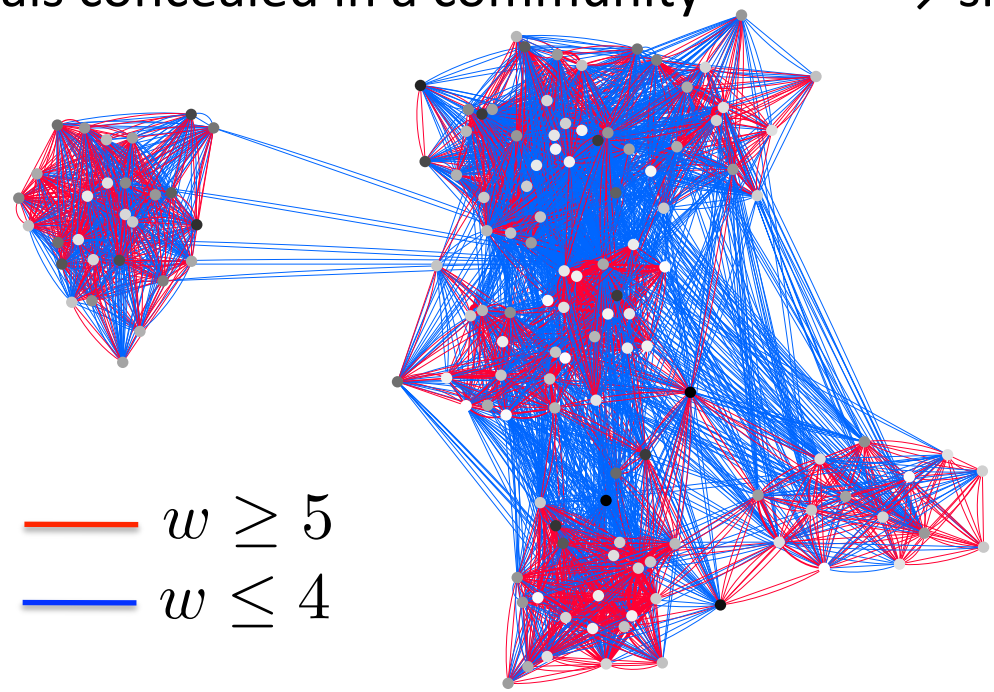
- Fix $k_i$, small $\overline{w}_i$ → abundance of weak links (with small weights)

☐ <u>Hypothesis about links</u>

- Weak links connect different communities.

☐ <u>Hypothesis about nodes</u>

- Individuals connecting different communities → large $I_i$
- Individuals concealed in a community → small



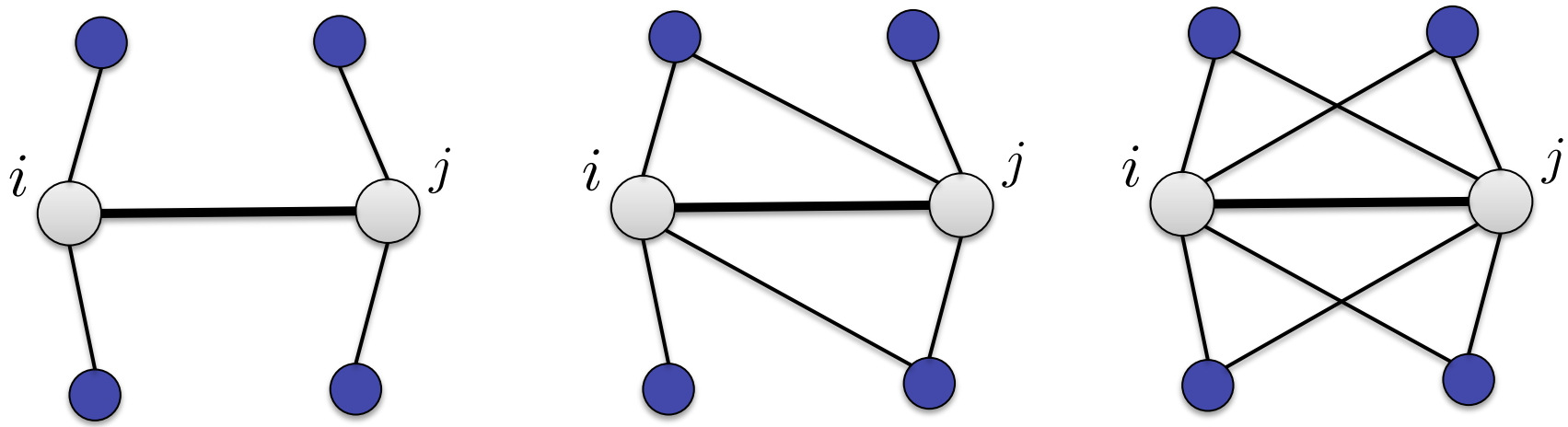
# Hypothesis about links

Purpose: quantify the degree of being concealed in a community

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

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

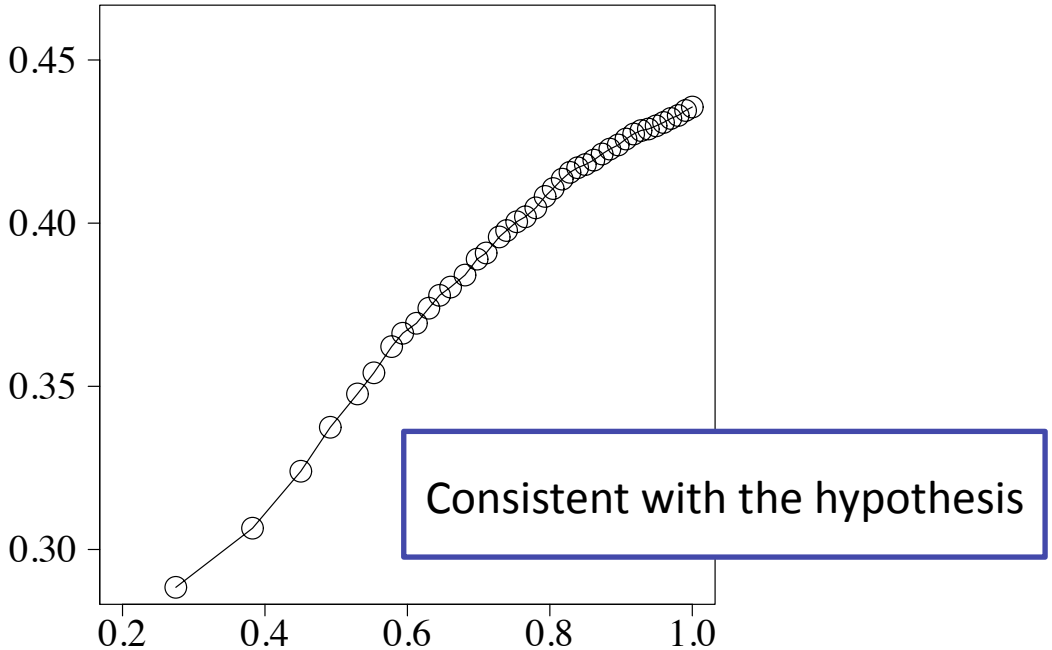
$O_{ij} = 0$ $\quad$ $O_{ij} = 1/2$ $\quad$ $O_{ij} = 1$



# Hypothesis about links

"Strength of weak ties" hypothesis;
a link with large $w_{ij}$ tends to have a large $O_{ij}$
→ $O_{ij}$ increases with $w_{ij}$

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




$P_{\mathrm{cum}}(w)$ fraction of the links with $w_{ij} \leq w$

# Hypotheses

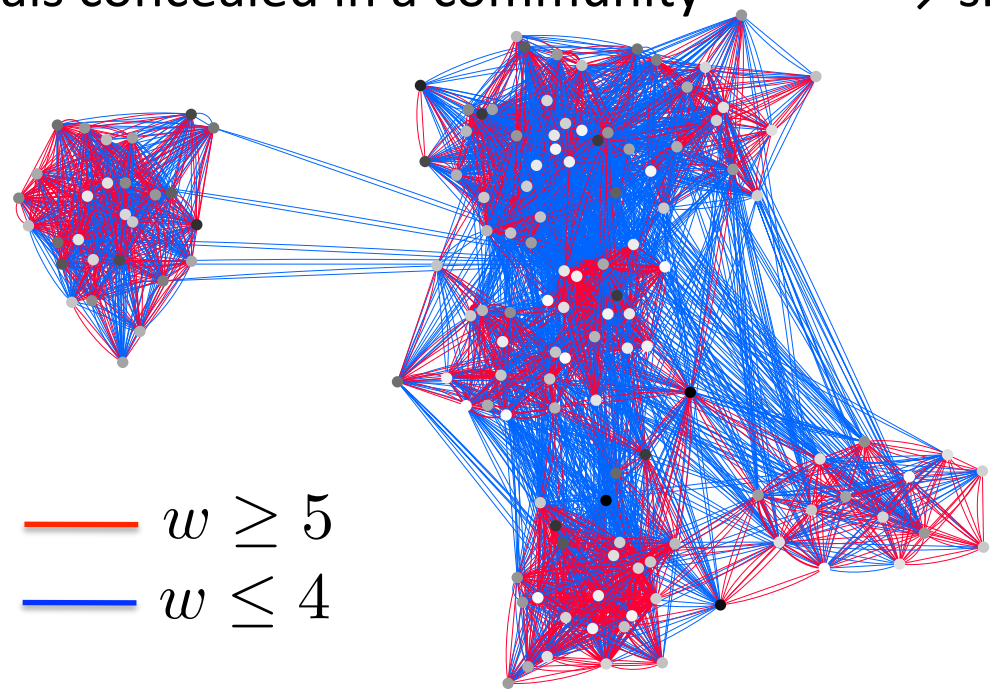
- Fix $k_i$, small $\overline{w}_i$ → abundance of weak links (with small weights)

☑ <u>Hypothesis about links</u>

- Weak links connects different communities.

☐ <u>Hypothesis about nodes</u>

- Individuals connecting different communities → large $I_i$
- Individuals concealed in a community → small



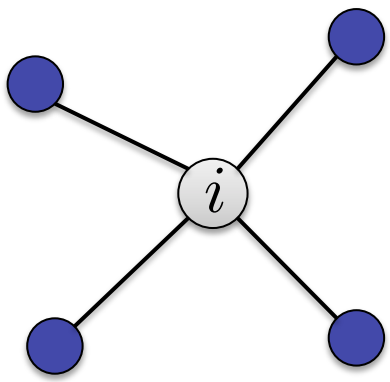
# Hypothesis about nodes

<u>Purpose</u>: quantify the degree of being concealed in a community

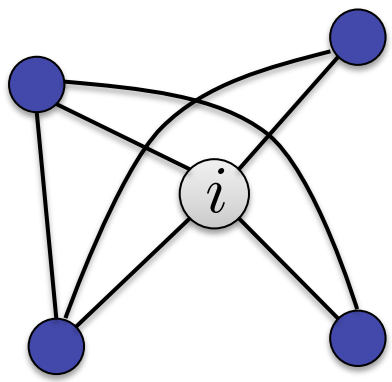
Clustering coefficient of a node (Watts and Strogatz, 1998)

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

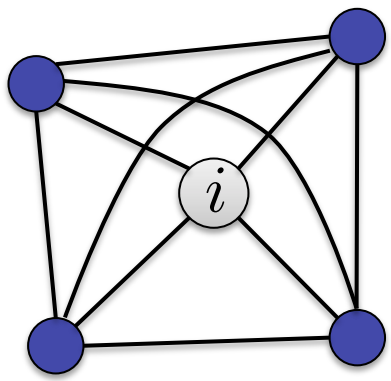
$C_i = 0$




$C_i = 1/2$

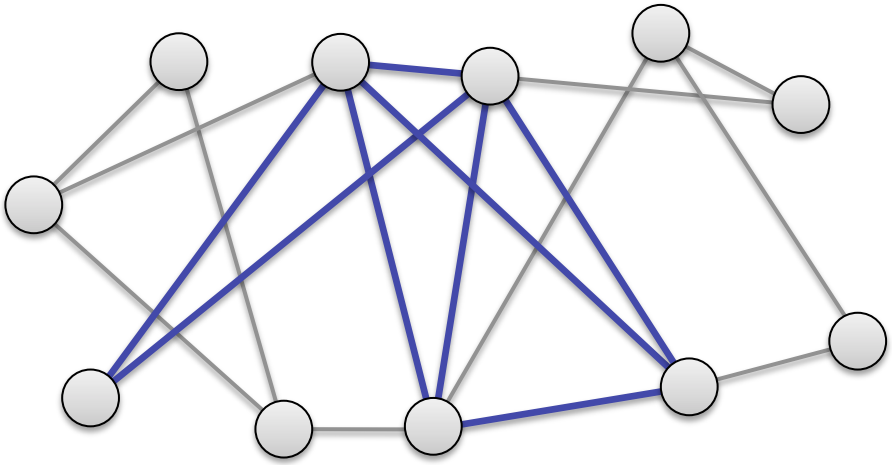



$C_i = 1$

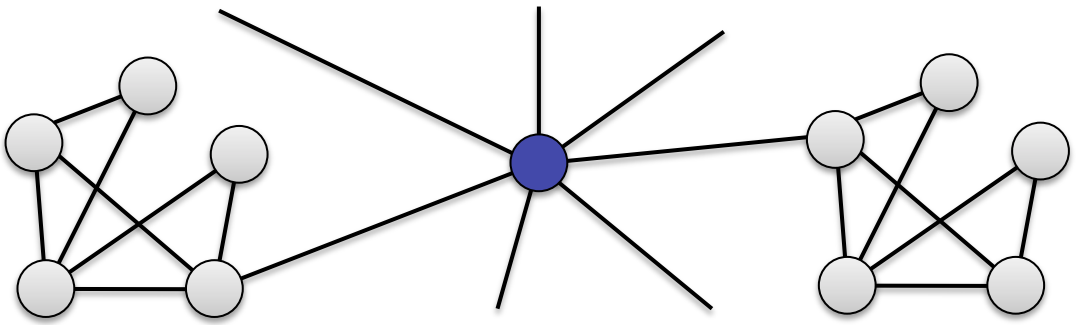


# Abundance of triangles

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



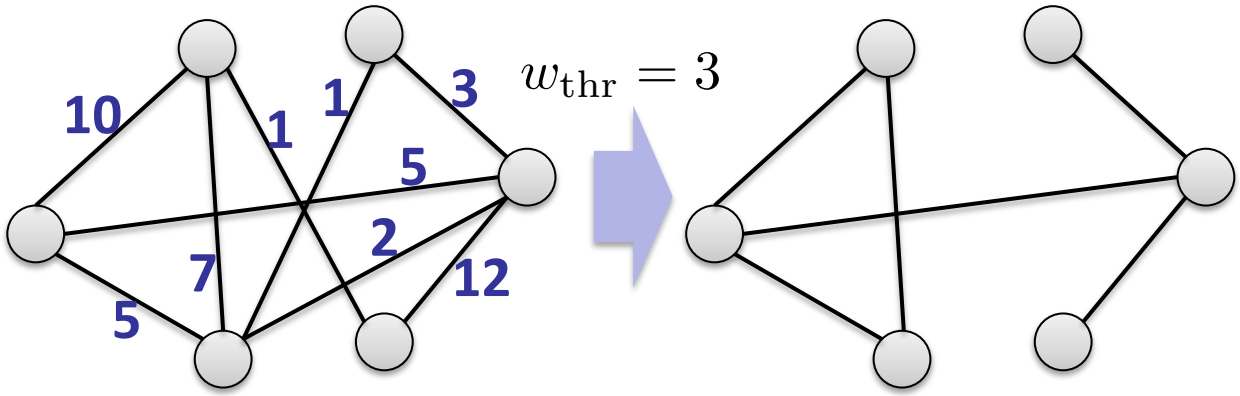
- Hierarchical structure (Ravász and Barabási, 2003)
  Nodes connecting communities → small $C_i$



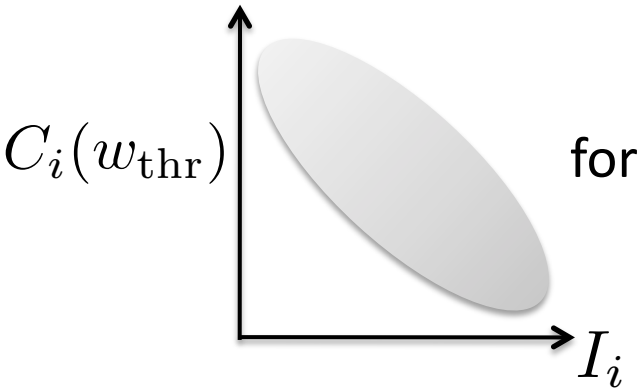
# Clustering coefficient with link-weight threshold

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

Original CN




<u>Expectation</u>: triangles persist around individuals with small $I_i$

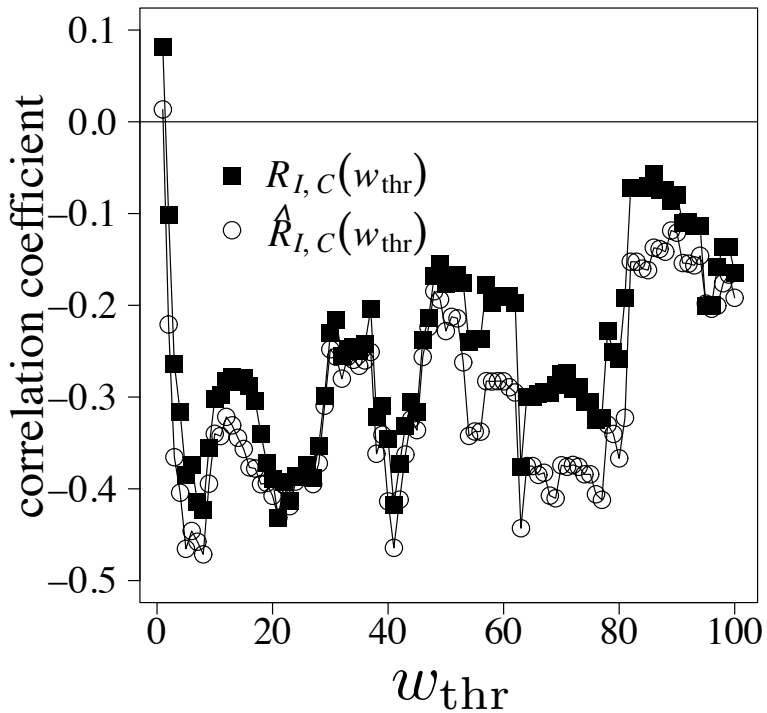



for a fixed $w_{\rm thr}$

# Hypothesis about nodes

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

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



# Hypotheses

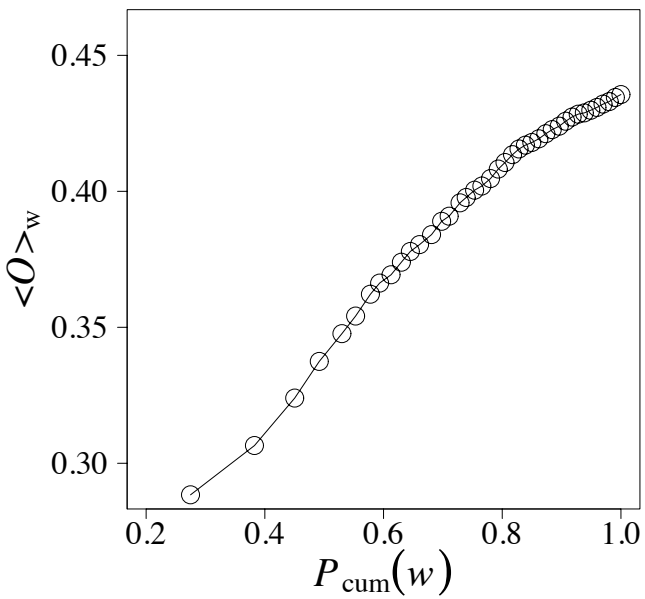
- Fix $k_i$, small $\overline{w}_i$ → abundance of weak links (with small weights)

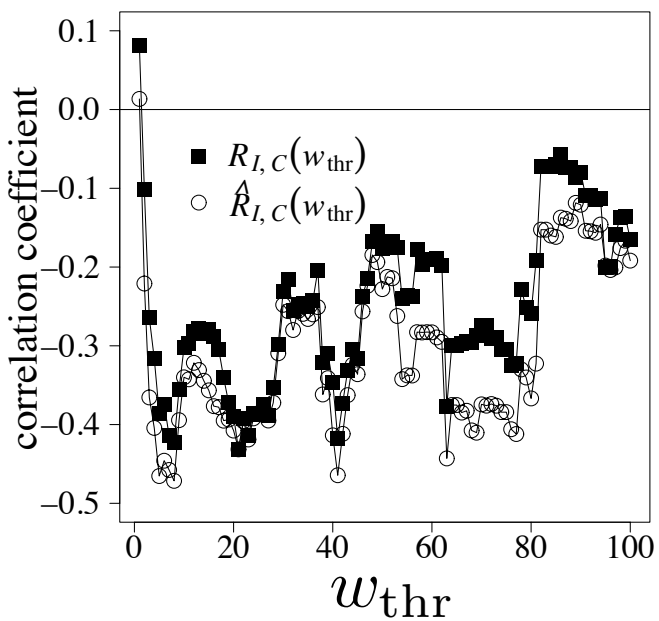
☑ <u>Hypothesis about links</u>

- Weak links connects different communities.

☑ <u>Hypothesis about nodes</u>

- Individuals connecting different communities → large $I_i$
- Individuals concealed in a community → small



# 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 → large $I_i$
  - INTRA-community individuals → 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