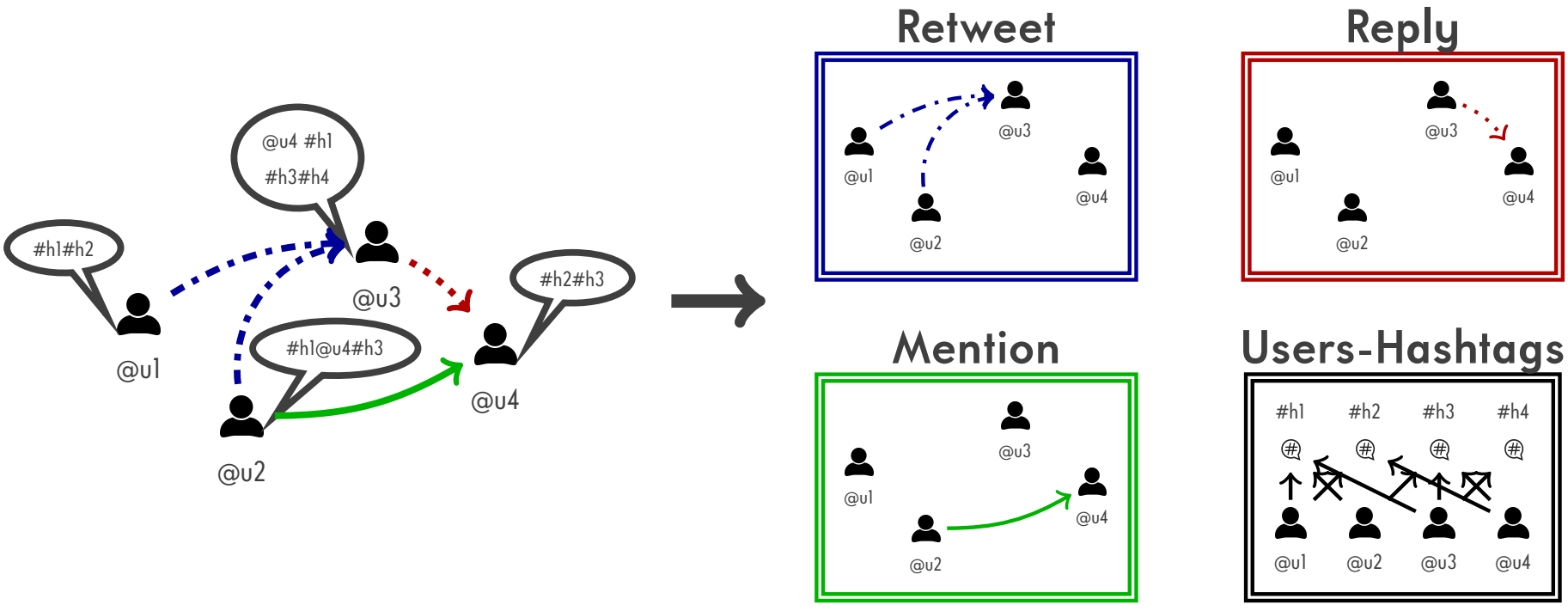
Tarragona, March 20, 2019

Link weights recovery in heterogeneous information networks

Hông-Lan Botterman[1], Robin Lamarche-Perrin[2]
[1]Sorbonne Université (LIP6)
[2]Institut des Systèmes Complexes Paris Île de France

Link weights recovery: an illustrative example



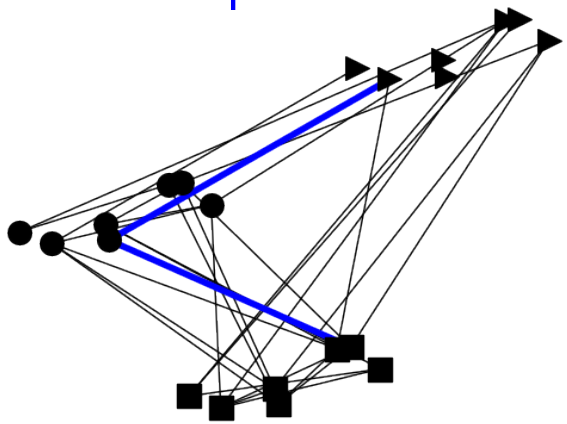
Expressing the UH link weights using other links/link weights

For some users, recovering their UH link weights, knowing some other links/link weights.

heterogeneous information network

$$H := (V, E, w, \mu_s, \mu_t, \mathcal{V}, \mathcal{E}, \phi, \psi)$$

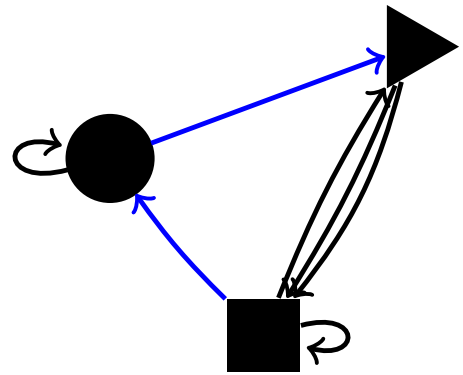
path



network schema

$$T_H := (\mathcal{V}, \mathcal{E}, \nu_s, \nu_t)$$

metapath: ■ → ● →▶



path-constrained random walk

$\mathcal{V} = \{V_1, \cdots, V_m\}$ and $\mathcal{E} = \{E_1, \cdots, E_r\}$

$$\mathcal{P} = V_1 \xrightarrow{E_{j_1}} V_{i_2} \ldots V_{i_{n-1}} \xrightarrow{E_{j_{n-1}}} V_n,\ i_2, \ldots, i_{n-1} \in \{1, \ldots, m\},$$
$$j_1, \ldots, j_{n-1} \in \{1, \ldots, r\}.$$

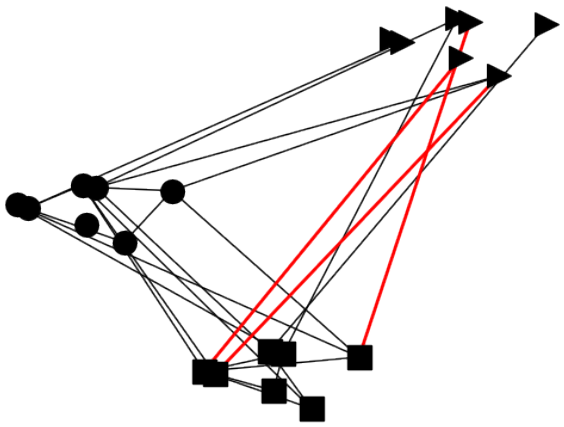
$$\mathbb{P}((v_n|v_1) \,|\, \mathcal{P}) = \sum_{v_{n-1} \in V_{i_{n-1}}} \frac{w_{E_{j_{n-1}}}(v_{n-1}, v_n)\, f^{\alpha}_{v_n}}{\sum_k w_{E_{j_{n-1}}}(v_{n-1}, v_k)\, f^{\alpha}_{v_k}} \mathbb{P}\left((v_{n-1}|v_1) \,|\, \mathcal{P}^{1,i_{n-1}}\right)$$

C. Shi, Y. Li, J. Zhang, Y. Sun, and P. S. Yu. 2017. A Survey of Heterogeneous Information Network Analysis. IEEE Trans. on Knowl. and Data

Linear combination of PCRW

Dependent variable: result w.r.t. a target link type - PCRW(E^*)

Independent variables: results of random walks w.r.t. particular link types - PCRW($\mathcal{P}$), $\mathcal{P} \in \mathcal{E}_\mathcal{P}$



$E^* = \blacksquare \rightarrow \blacktriangleright$: target metapath

$$\left.\begin{array}{l} \mathcal{P}_1 = \blacksquare \rightarrow \blacksquare \rightarrow \blacktriangleright \\ \mathcal{P}_2 = \blacksquare \rightarrow \bullet \rightarrow \blacktriangleright \\ \mathcal{P}_3 = \blacksquare \rightarrow \bullet \rightarrow \bullet \rightarrow \blacktriangleright \end{array}\right\} \mathcal{E}_\mathcal{P}$$

$$\underbrace{F((\blacksquare, \blacktriangleright); \mathcal{E}_\mathcal{P})}_{\text{approx. value}} := \beta_0 + \sum_{\mathcal{P} \in \mathcal{E}_\mathcal{P}} \beta_\mathcal{P} \, \text{PCRW}(\mathcal{P})$$

$$\sim \text{PCRW}(E^*)$$

PCRW($\cdot$): path-constrained random walk wrt $\cdot$

Selection by a forward linear regression

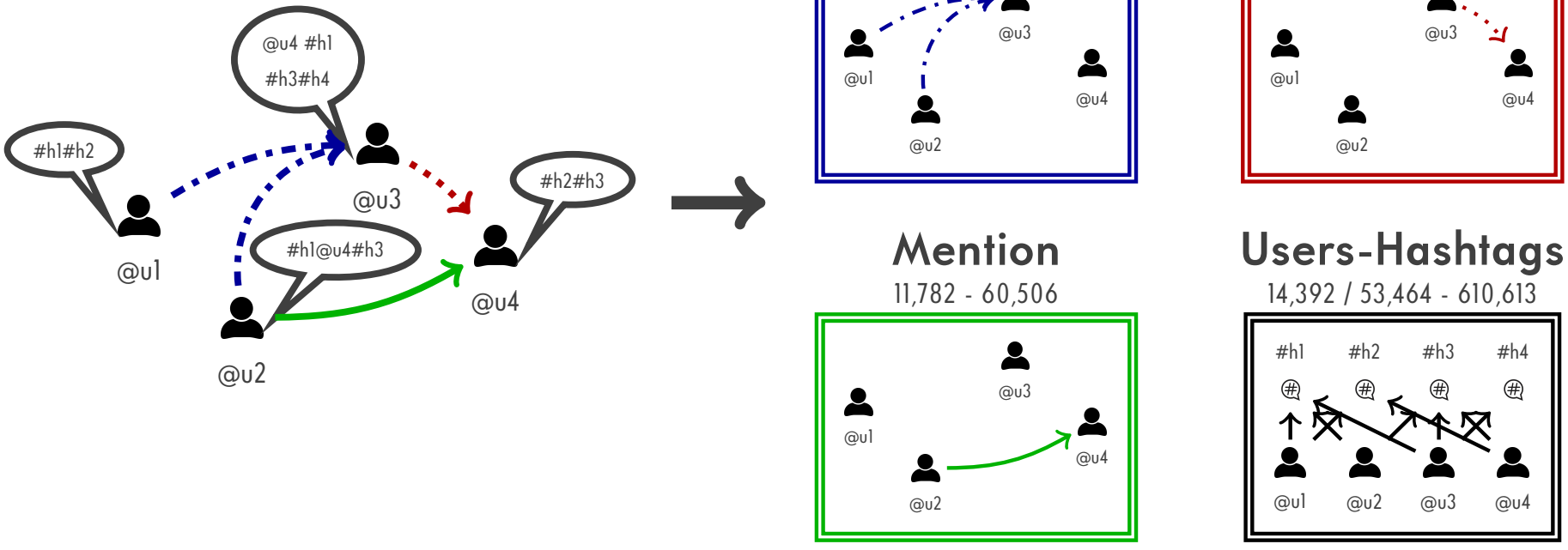
1. **Least squares problem:**

   Let $S = \sum r_i^2$: $\quad \dfrac{\partial S}{\partial \beta_j} = 2 \sum r_i \dfrac{\partial r_i}{\partial \beta_j} = 0$

   $\longrightarrow \hat{\beta} = (X^T X)^{-1} X^T y$

2. **Forward linear regression:**
   - starting with no variables in the model
   - testing the addition of each variable using a chosen model fit criterion: p-value and t-test: $t^* = (\bar{x} - \mu_0)/(s/\sqrt{n})$
   - adding the variable (if any) whose inclusion gives the most statistically significant improvement of the fit:
     cœf. of determination $R^2 = 1 - \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \; / \; \sum_{i=1}^{n} (y_i - \bar{y_i})^2$
   - repeating this process until there is no further improvement

Twitter data related to FIFA World Cup 2014



From June 12 to July 13, 2014
32 teams - 64 matches played

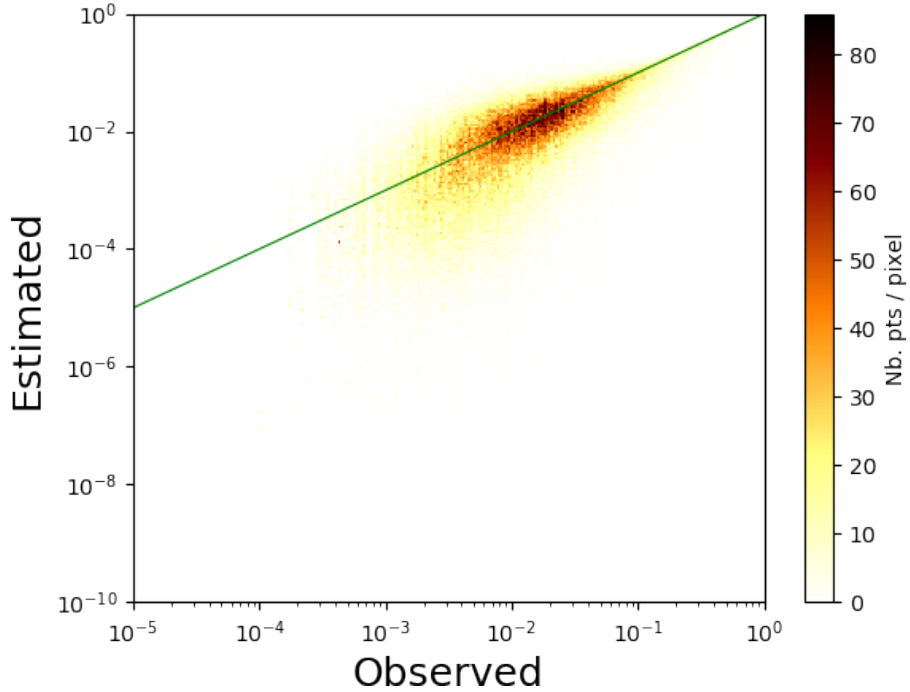
HIN with $\mathcal{V}$ = {users, hashtags} and $\mathcal{E}$ = {retweet, reply, mention, post}.

Describing UH from other link types (locally)

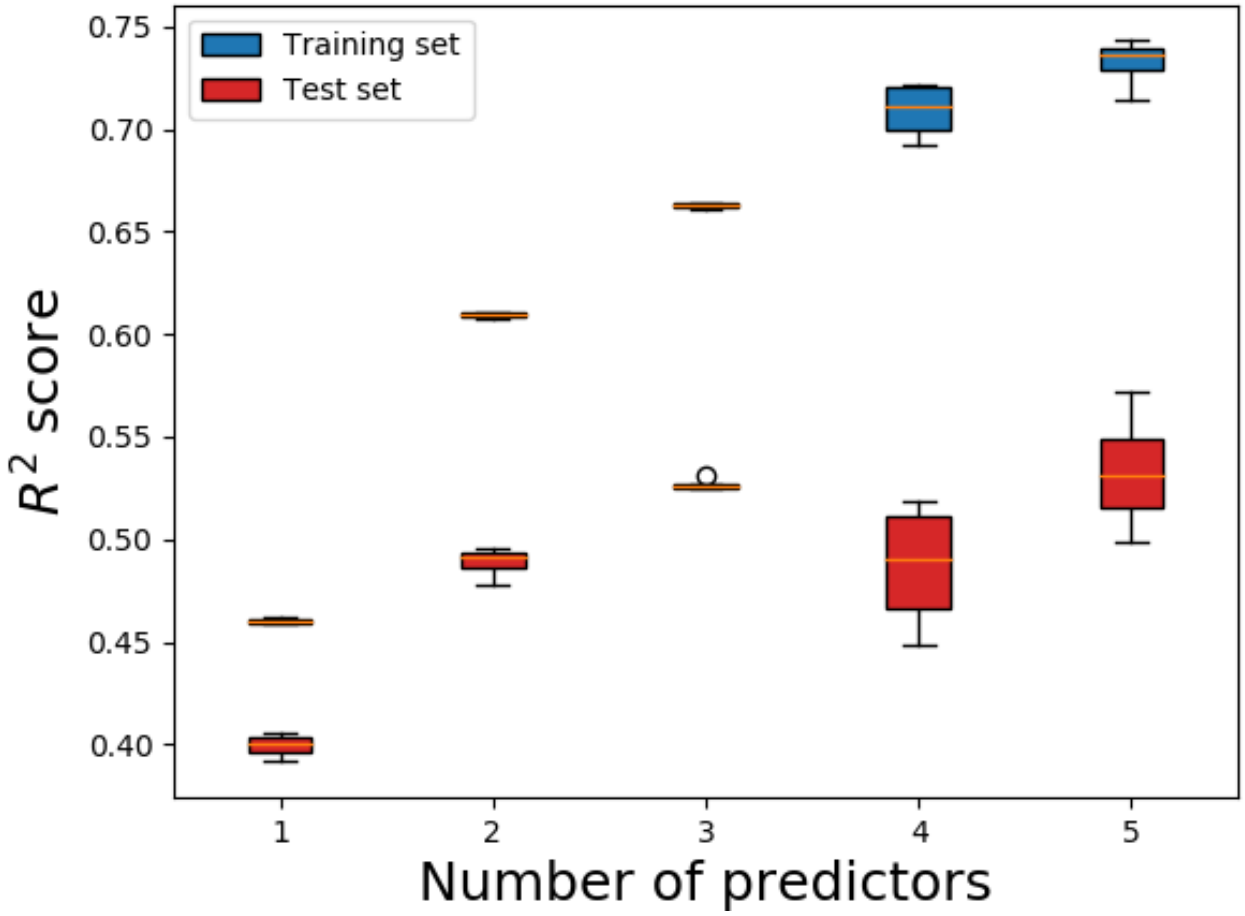
$\mathcal{E}_{\mathcal{P}}$ = { Metapaths of length ≤ 4 }

UH = 0.1974 MT-UH (iter 1)
+ 0.5556 RP-UH (iter 2)
+ 0.0650 RT-RP-UH (iter 3)
+ 0.1591 RP-MT-UH (iter 4)
+ 0.0074 MT-RT-UH (iter 5)

with $r^2 = 0.7129$



Recovering UH from other link types and other users



Training set 80% - Test set: 20% of users

Overfitting

Summary

Main idea

Adequate linear combination of path-constrained random walks results to describe and, to some extent, retrieve the strength of the links between different entities.

Several improvements

- Temporal aspects
- Searching for candidate metapaths

Tarragona, March 20, 2019

Link weights recovery in heterogeneous information networks

Hông-Lan Botterman[1], Robin Lamarche-Perrin[2]
[1]Sorbonne Université (LIP6)
[2]Institut des Systèmes Complexes Paris Île de France