Combining path-constrained random walks to recover link weights in heterogeneous information networks

Hông-Lan Botterman, Robin Lamarche-Perrin

Combining path-constrained random walks to recover link weights in heterogeneous information networks

Hông-Lan Botterman[1] and Robin Lamarche-Perrin[2]

[1]Sorbonne Université, CNRS, LIP6, F-75005 Paris, France
[2]CNRS, Institut des système complexes de Paris Île-de-France, ISC-PIF, UPS 3611, Paris, France

Abstract

Heterogeneous information networks (HIN) are abstract representations of systems composed of multiple types of entities and their relations. Given a pair of nodes in a HIN, this work aims at recovering the exact weight of the incident link to these two nodes, knowing some other links present in the HIN. Actually, this weight is approximated by a linear combination of probabilities, results of path-constrained random walks i.e., random walks where the walker is forced to follow only a specific sequence of node types and edge types which is commonly called a meta path, performed on the HIN. This method is general enough to compute the link weight between any types of nodes. Experiments on Twitter data show the applicability of the method.

1. Introduction. Networked entities are ubiquitous in real-world applications. Examples of such entities are humans in social or communication activities and proteins in biochemical interactions. Heterogeneous information networks (HIN), abstract representations of systems composed of multiple types of entities and their relations, are good candidates to model such entities, together with their relations, since they can effectively fuse a huge quantity of information and contain rich semantics in nodes and links. In the last decade, the heterogeneous information network analysis has attracted a growing interest and many novel data mining tasks have been designed in such networks, such as similarity search, link prediction, clustering and classification just to name a few.

The goal of this work is to recover, for a given pair of nodes in a weighted HIN, the actual incident link weight to these two nodes, knowing some other links present in the HIN. Trying to capture not only the presence of a link but also its actual weight can be useful, for instance, in recommendation systems where the weight can be taken for the “rating" a user would give to an item. Another application would be the detection of disease-gene candidate thanks to the prediction of protein-protein interactions.

This problem can be related to the node similarity problem since similar nodes tend to be connected. Indeed, the similarity score between two nodes, result of a particular function of these two nodes, can be seen as the strength of their connection and hence, the link weight connecting them. Here, the particular function is related to a random walk on the graph.

In HIN, most of similarity scores [6, 9] are based on the concept of meta path, roughly defined as a concatenation of node types linked by corresponding link types. The type of a node/link is basically a label in the abstract representation. Meta paths can be used as a constraint to a classic random walk: the walker is allowed to take only paths satisfying a particular meta path. These path-constrained random walks have the sensitivity to take into account explicitly different semantics present in HIN.

Back to our goal, the target weight is approximated by a linear combination of probabilities, results of path-constrained random walks performed on the HIN. The proposed method aims at finding a relevant set of meta paths and the best possible cœfficients such that the difference between the exact link weight and its approximation is minimized.

The rest of this paper is organized as follows. In Section 2, we remind basic concepts about HIN and present the problem statement. Section 3 explains our method and we apply it on Twitter data related to

the Football World Cup 2014 in Section 4. We review some related work in Section 5 and finally, we conclude and give some perspectives in Section 6.

2. Preliminary Concepts. In this section, we remind some basic concepts of weigthed HIN useful for the following and define the "weight recovering" problem. Fig. 1 illustrates this section.

Definition 2.1 (Weighted directed multigraph) *A weighted directed mutligraph is a 5-tuple* $G := (V, E, w, \mu_s, \mu_t)$ *with* V *the node set,* E *the link set,* $w : E \to \mathbb{R}$ *the function that assigns each link a real weight,* $\mu_s : E \to V$ *the function that assigns each link a source node,* $\mu_t : E \to V$ *the function that assigns each link a target node.*

Definition 2.2 (Heterogeneous Information Network) *A HIN* $H := (G, \mathscr{V}, \mathscr{E}, \phi, \psi)$ *is a weighted directed multigraph* G *along with* $\mathscr{E}$ *the node type set,* $\mathscr{V}$ *the link type set,* $\phi : E \to \mathscr{E}$ *the function that assigns a node type to each node and* $\psi : V \to \mathscr{V}$ *the function that assigns a link type to each link such that if two links belong to the same link type, the two links share the same starting and target node type i.e.,* $\forall e_1, e_2 \in E$, $\big(\psi(e_1) = \psi(e_2)\big) \Rightarrow \big(\phi(\mu_s(e_1)) = \phi(\mu_s(e_2)) \wedge \phi(\mu_t(e_1)) = \phi(\mu_t(e_2))\big)$.

Understanding the node types and link types in a complex HIN is not always easy, thus it is sometimes necessary to provide the meta level (i.e., schema-level) description of the network. Therefore, the concept of network schema is proposed to describe the meta structure of a network.

Definition 2.3 (HIN Schema) *Let* H *be a HIN. The schema* T_H *for* H *is a directed graph defined on the node types* $\mathscr{V}$ *and link types* $\mathscr{E}$ *i.e.,* $T_H := (\mathscr{V}, \mathscr{E}, \nu_s, \nu_t)$ *with* $\nu_s : \mathscr{E} \to \mathscr{V} : E^* \mapsto \nu_s(E^*) := \phi\big(\mu_s(e)\big)$ *the function that assigns each link a source node and* $\nu_t : \mathscr{E} \to \mathscr{V} : E^* \mapsto \nu_t(E^*) := \phi\big(\mu_t(e)\big)$ *the function that assigns each link a target node, where* $e \in \psi^{-1}(E^*)$ *and* ψ^{-1} *the pseudo-inverse of* ψ *defined by* $\psi^{-1} : \mathscr{E} \to 2^E : E^* \mapsto \{e \in E \,|\, \psi(e) = E^*\}$.

We can effectively take any element $e \in \psi^{-1}(E^*)$ since $\{e \in E \,|\, \psi(e) = E^*\}$ is the equivalence class of any of its elements, with the equivalence relation "has the same type of". By definition of HIN, it is sufficient to take one member of the equivalence class to know the node types the link type connects.

Two entities in a HIN can be linked via different paths and these paths have different semantics. These paths can be defined as meta paths as follows.

Definition 2.4 (Meta path) *A meta path* $\mathscr{P}$ *of length* $n-1 \in \mathbb{N}$ *is a sequence of node types* $V_1, \cdots, V_n \in \mathscr{V}$ *linked by link types* $E_1, \cdots, E_{n-1} \in \mathscr{E}$ *as follows:* $\mathscr{P} = V_1 \xrightarrow{E_1} V_2 \cdots V_n \xrightarrow{E_{n-1}} V_n$ *which can also be denoted as* $\mathscr{P} = E_1 E_2 \cdots E_{n-1}$.

Given a meta path $\mathscr{P} = V_1 \xrightarrow{E_1} V_2 \cdots V_n \xrightarrow{E_{n-1}} V_n$ and a path $P = v_1 \xrightarrow{e_1} v_2 \cdots v_{n-1} \xrightarrow{e_{n-1}} v_n$, if $\forall i \in \{1, ..., n\}$, $\phi(v_i) = V_i$, $\forall i \in \{1, ..., n-1\}$, $\mu_s(e_i) = v_i$, $\mu_t(e_i) = v_{i+1}$ and $\psi(e_i) = E_i$, then the path P satisfies the meta path $\mathscr{P}$ and we note $P \in \mathscr{P}$. Hence, a meta path is a set of paths.

Problem 2.1 (Combination of meta paths) *Let be a HIN* $H = (G, \mathscr{V}, \mathscr{E}, \phi, \psi)$, *with* $G = (V, E, w, \mu_s, \mu_t)$ *a directed weighted multigraph, and a target meta path* E_c *between two node types. The problem is to find a set of relevant meta paths* $\mathscr{E}_{\mathscr{P}}$ *and a linear function* F *of (functions that themselves depend on) these meta paths that best quantifies, for each pair of nodes in* H*, the strength of their connection via* E_c.

3. Method. We present our method for solving Problem 2.1 in three steps.

Without loss of generality, $\mathscr{V} = \{V_1, \cdots, V_m\}$, $\mathscr{E} = \{E_1, \cdots, E_r\}$ and we note E_c the target meta path defined between V_1 and V_n. We consider a meta path $\mathscr{P} = V_1 \xrightarrow{E_{j_1}} V_{i_2} \cdots V_{i_{n-1}} \xrightarrow{E_{j_{n-1}}} V_n$ different from E_c where $i_2, \cdots, i_{n-1}$ and $j_1, \cdots, j_{n-1}$ are all indices that can take integer values between 1 and m and 1 and r respectively.

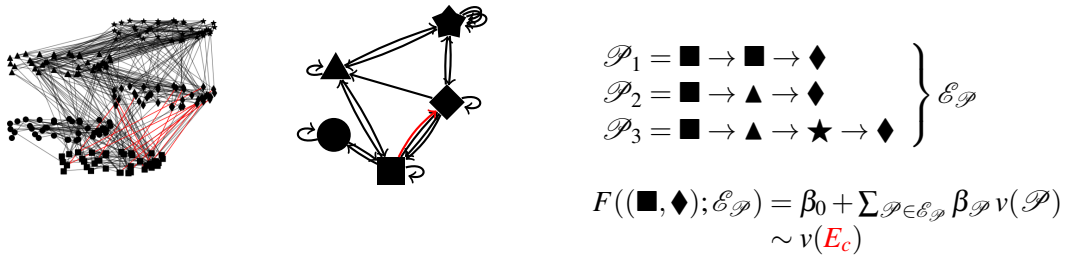

Figure 1: (Left) Example of HIN composed of five node types, represented by diverse shapes, and multiple link types. (Middle) Its associated network schema composed by five nodes and twenty links. (Right) Illustration of the problem statement. For each pair of nodes in (■,♦), there is possibly a path connecting them. The link weight is approximated by a linear combination of the path-constrained random walk results.

3.1. Path-Constrained Random Walk. Given $v_n \in V_n$ and $v_1 \in V_1$, the probability of reaching v_n from v_1 following the meta path $\mathscr{P}$ is simply defined by the random walk starting at v_1 and ending at v_n following only paths satisfying $\mathscr{P}$. Formally,

$$\begin{aligned}\mathbb{P}((v_n|v_1)\,|\,\mathscr{P}) &= \sum_{v_{n-1}\in V_{i_{n-1}}} \mathbb{P}\Big((v_n|v_{n-1})\,|\,\mathscr{P}^{i_{n-1},n}\Big)\,\mathbb{P}\Big((v_{n-1}|v_1)\,|\,\mathscr{P}^{1,i_{n-1}}\Big) \\ &= \sum_{v_{n-1}\in V_{i_{n-1}}} \frac{w_{E_{j_{n-1}}}(v_{n-1},v_n)}{\sum_k w_{E_{j_{n-1}}}(v_{n-1},v_k)}\,\mathbb{P}\Big((v_{n-1}|v_1)\,|\,\mathscr{P}^{1,i_{n-1}}\Big) \qquad (1)\end{aligned}$$

with $\mathscr{P} =: \mathscr{P}^{1,n}$, $\mathscr{P}^{a,b}$ the truncated meta path of $\mathscr{P}$ from node type V_a to V_b, $V_a, V_b \in \mathscr{V}$, $w_{E_i}(v_j,v_k)$ the link's weight of type E_i between nodes v_j and v_k and $\mathbb{P}((v_2|v_1)|\mathscr{P}^{1,i_2}) = w_{E_{j_1}}(v_1,v_2)/\sum_k w_{E_{j_1}}(v_1,v_k)$ the basis of recurrence.

Furthermore, we forbid the walker to return to the initial node on the penultimate step of the walk i.e., if $V_{i_{n-1}} = V_1$, the sum in eq. (1) only holds for all $v_{n-1} \neq v_1$. It prevents us from using what we are looking for to find what we are looking for.

3.2. Linear Regression Model. Since H is a HIN, multiple types of links can connect the nodes. Hence, there is no reason to restrict ourselves to a single meta path to compute the reachability of one node from another. As a result, the similarity between v_n and v_1 is defined by several path-constrained random walk results combined through a linear regression model of the form

$$F((v_n|v_1)\,|\,\mathscr{E}_{\mathscr{P}}) = \beta_0 + \sum_{\mathscr{P}\in\mathscr{E}_{\mathscr{P}}} \beta_{\mathscr{P}}\,\mathbb{P}((v_n|v_1)\,|\,\mathscr{P})$$

where $\mathscr{E}_{\mathscr{P}}$ is the set of relevant meta paths and the vector $\boldsymbol{\beta} := [\beta_0,\beta_1,\cdots,\beta_{|\mathscr{E}_{\mathscr{P}}|}]^T$ is real-valued cœfficients. The cœfficients stress the contribution of each meta path in the final similarity score i.e., our approximation $F((v_n|v_1)\,|\,\mathscr{E}_{\mathscr{P}})$ of the exact link weight $w_{E_c}(v_1,v_n)$. The choice of linear model is simply motivated by its interpretation in our particular case. Since the components of $\boldsymbol{\beta}$ are not confined in [0,1] and do not sum to 1, F is a real-valued function whose image is neither confined in [0,1]. Given example node pairs and their link weights, $\boldsymbol{\beta}$ is estimated by the least squares method which is appreciated for its applicability and simplicity.

3.3. Forward Selection Procedure. In order to determine the set $\mathscr{E}_{\mathscr{P}}$, we use the forward selection with p-value and r^2 criteria. This is a greedy approach but very simple and intuitive. The p-values are used to test the significance of each predictor. Given the hypothesis $H_0 : \beta = 0$ against the hypothesis $H_1 : \beta \neq 0$, the p-value p is the probability, under H_0, of getting a statistics as extreme as the observed

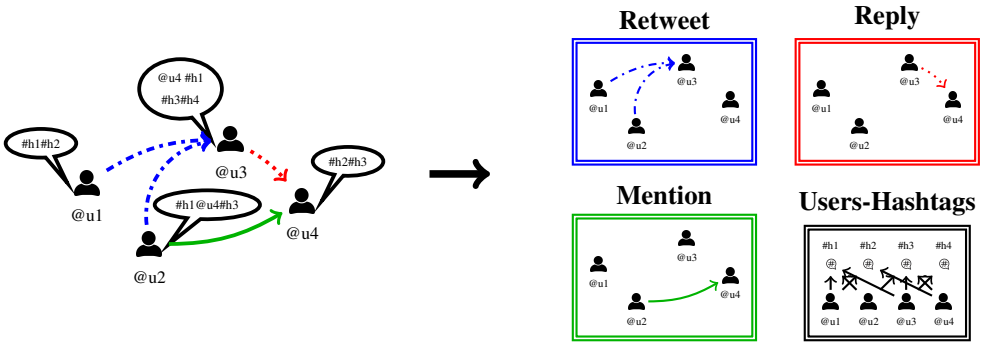

Figure 2: Illustration of the graphs construction representing the Twitter interactions. The underlying HIN is such that $\mathscr{V}$ ={users, hashtags} and $\mathscr{E}$ ={RT, RP, MT, UH}.

value on the sample. We reject the hypothesis H_0, at the level α, if $p \leq \alpha$ in favor of H_1. Otherwise, we reject H_1 in favor of H_0. Conversely, the r^2 score is used to test the quality of the entire model. It is the proportion of the variance in the dependent variable that is predictable from the predictors.

So, given k predictors or explanatory variables which are the meta paths, the forward selection procedure works as follows

- Start with a null model i.e. no predictor but only an intercept. Typically, this is the average of the dependent variable;
- Try k linear regression models and chose the one which gives the best model with respect to the criterion. In our case, the one that maximizes the coefficient of determination r^2;
- Search among the remaining variables the one that, added to the model, gives the best result i.e., the higher r^2 such that all the variables in the model are significant i.e., their p-value is below the chosen threshold. Iterate this step until no further improvement.

4. Experiments. We present the dataset on which we test the proposed method as well as the construction of the resulting graphs. Then, we report our results concerning different tests namely, the importance of meta path length, a description task and finally a recovery task.

4.1. Dataset Description and Setup. The data we use is a set of tweets collected from Twitter during the Football World Cup 2014. This period extents from June 12 to July 13, 2014. Twitter allows multiple kinds of interactions between its users. Here, we consider retweet (RT), reply (RP) and mention (MT) actions plus the fact of posting hashtags (UH).

Based on these actions, we construct a HIN with node types $\mathscr{V}$ ={users, hashtags} and edge types $\mathscr{E}$ ={RT, RP, MT, UH} as illustrated in Fig. 2. Each node represents a user or a hashtag. We create a link from u_1 to u_2 if u_1 retweets, replies (to) or mentions u_2 and the weight of the link correspond to the number of times u_1 performs the specific action towards u_2 during the whole world cup. For the user-hashtag graph, a link exists between u and h if h appears in u's post and the weight of the link corresponds to the number of times u post h during the whole world cup.

The RT graph is composed of 6069 nodes and 19495 links, the RP graph is composed of 8560 nodes and 11782 links and the MT graph is composed of 11782 nodes and 60506 links. The Pearson coefficient between the stochastic matrices rises to 0.1776, 0.6783 and 0.4286 for RT/RP, RT/MT et RP/MT respectively. Thus, the retweet and mention relationships are clearly correlated which may cause some problems for the proposed method, as we shall see, since it is well known that least squares method is sensitive to that. Since the data is related to the world cup, the most used hashtags of bipartite users-hashtags graph are those referring to the 32 countries involved in the final phase as well as those referring directly to the event (#WorldCup2014, #Brazil, #Brasil2014, #CM2014, ...). The semi finalists have the greatest in-strength.

4.2. Results. We apply the proposed method to find if the hashtags posted by users (UH) can be explained by other relations (RT, RP, MT and their combinations). For instance, given a user u, explaining UH by RT-UH and MT-RP-UH means that the hashtags posted by u are, to some extent, a combination of those posted by the users retweeted by u and those posted by the users who received a response from users mentioned by u. In other words, we try to understand if, in the case of the football World Cup 2014, the probability that users post hashtags can be explained by the relations these users have with other users and the probability that these latter have to post specific words.

4.2.1. Meta Paths of Length 2. We test linear regression models with all the possible combinations of variables of length 2 (see Table 1). This test allows a first glimpse at the contribution of the simplest predictors. First, the more the predictors, the better is the value of the r^2. Nevertheless, it does not mean that all variables are significant. Indeed, the analysis of the coefficients and p-values makes it possible to realize the correlation of some variables. In models 5 and 7, the RT-UH cœfficient is negative with p-value greater than 0.05, consequence of the correlation with the MT-UH variable.

In summary and according to Table 1, the best model would be the model 4 whose predictors are RT-UH and RP-UH. This means that, for a given user, the hashtags she posts can be explained by the hashtags posted by the users she retweets with a contribution of 0.5795 and the users she replies to with a contribution of 0.3957. This model accounts for 61.16% of the variance.

Mod.	Var.	Cœf.	p-values	r^2
0	Average : 1.8704e-05			0.2992
1	RT-UH	0.6273	-	0.3594
2	RP-UH	0.4291	-	0.2289
3	MT-UH	1.0289	-	0.4606
4	RT-UH RP-UH	0.5795 0.3957	0.0062 0.0105	0.6116
5	RT-UH MT-UH	-0.3578 1.4534	0.0612 0.0087	0.5943
6	RP-UH MT-UH	0.0051 0.9391	0.0138 0.0057	0.6111
7	RP-UH RP-UH MT-UH	-0.1283 0.0791 1.1466	0.0791 0.0113 0.0111	0.6818

Table 1: Cœfficients and p-values for linear regressions whose variables correspond to meta paths of length 2. Model 0 corresponds to the null model: no predictor but one intercept that is the average of the explained variable.

4.2.2. Importance of Meta Path Length. This subsection looks at the length of the meta paths for a given link type. More specifically, we compute, for each link type, the r^2 score when the only predictor is associated to a random walk of length $l = 1, ..., 10$ in the same link type. Intuitively, the importance of a meta path decreases with its length since considering longer meta paths means considering neighborhoods more extended, hence the information is more diffused. This is corroborated with the left panel of Fig.3. Each link type brings a different quantity of information and the MT type is the more informative for our purpose. Plus, this test exposes a characteristic of the reply dynamics: most of the time, the replies involved only two people [7]. This is reflected through the oscillations of the reply scores. The scores associated to odd length random walks are low since the walker is forbidden to return to the initial node on the penultimate step of the walk. We also draw in black the scores when we do not differentiate the link types i.e., all the link weights between to nodes are aggregated. This score is below the average score of the three link types. One can see that just take the mention or retweet type is more informative than the aggregation.

The right panel of Fig. 3 shows r^2 scores when we combine variables of different lengths related to the same link type in the model. Actually, the score associated to the abscissa l is related to the model whose predictors are meta paths of length smaller or equals to $l + 1$ and whose the l first steps are in the same type of links. Again, the more the variables, the better the score. Also, the increase is not linear; the best improvement happens when we combine length-1 and length-2 variables. We can also observe that scores given by the RT and MT types are really similar when considering more than two variables while

there is a clear difference in the r^2 score for single variable. Once again, the score for the aggregation is shown and is far below the other scores. This indicates that it is potentially interesting to distinguish the types of links.

In summary, these tests tend to show that considering too long as well as too many meta paths is not necessarily useful in our case.

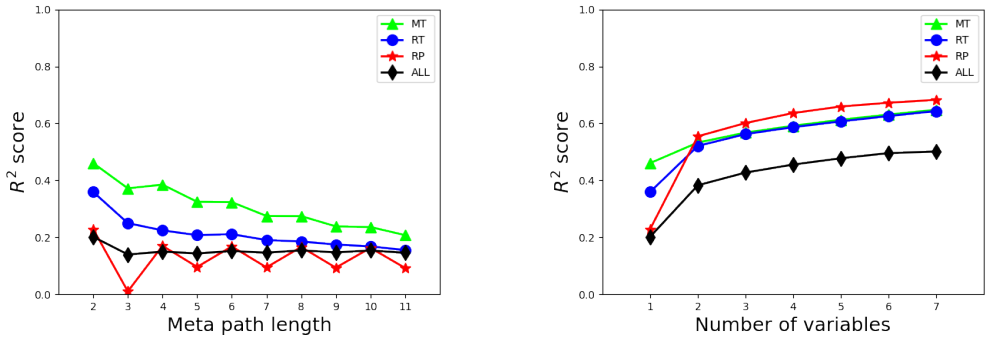

Figure 3: (Left) Linear regression r^2 scores with one meta path according to its length. (Right) Linear regression r^2 scores according to the number of meta paths of the same link type.

4.2.3. Forward Linear Regression for Description. We apply the proposed algorithm on the entire dataset with a threshold $\alpha = 0.05$ for p-values. The number of meta paths grows exponentially with the length and since the length is unbounded, the set of possible meta paths is infinite. Here, the k potential predictors are those of length less than or equal to 4. This is motivated by the test performed in the previous subsection. In addition, the semantics of longer paths are less clear than shorter paths.

Results are reported in Table 2. The final model contains five predictors related to meta path whose length are no longer than 3 and no intercept. This regression model accounts for 71.29% of the variance. To comfort the goodness of fit of the model, we plot in Fig. 4 the density plot in log-log scale of the observed values versus the estimated values. The green line represents the ideal case where estimated values match observed ones. Most of the data points fall to this line which indicates that linear model is a good choice.

Mod.	Var.	Cœf.	p-values	r^2
0	Average: 1.8704e-05			0.2992
1	MT-UH	1.0289	-	0.4606
2	MT-UH	0.9391	0.0057	0.6112
	RP-UH	0.0052	0.0137	
3	MT-UH	0.8464	0.0062	0.6682
	RP-UH	0.0335	0.0124	
	RT-RP-UH	0.1077	0.0138	
4	MT-UH	0.8114	0.0063	0.6947
	RP-UH	0.0362	0.0109	
	RT-RP-UH	0.0766	0.0142	
	RP-MT-UH	0.0676	0.0143	
5	MT-UH	0.1974	0.0094	0.7129
	RP-UH	0.5556	0.0146	
	RT-RP-UH	0.0650	0.0125	
	RP-MT-UH	0.1591	0.0160	
	MT-RT-UH	0.0074	0.0124	

Table 2: Results for the forward stepwise linear regression.

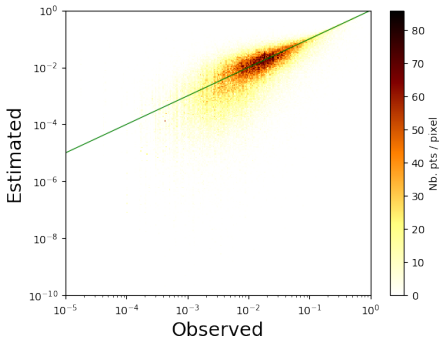

Figure 4: Density plot of observed versus estimated values for the model 5. Green line represents the perfect matching between observed and estimated data.

The best improvement comes with the addition of the second variable. The model with two predictors

is actually a local extremum (see Table 1). This allows to point two weaknesses of the method: there is no guarantee of finding the best model and the order of the variables selection is important. Note that the first two variables are part of the most direct relationships (meta paths of length 2) which is intuitive: the direct neighborhood of a user thus created shares common topics with her. The last meta path included in the model provokes an important change in the other cœfficients. This suggests this meta path is either correlated to other meta paths already present in the model or the presence of outliers. It is well known that ordinary least squares method is sensitive to that.

4.2.4. Forward Linear Regression for Recovery Task. We validate the method by performing a task aiming to recover the weights of missing links. In other words, this part tries to answer to the question: is it possible to know, in a quantitative way, the way some people post some hashtags, knowing the functioning of some other people ? To do so, we select 80% of the users and train the algorithm on it to obtain the vector $\boldsymbol{\beta}$. Then, we use it on the remaining 20% and compute the r^2 associated to each model.

Since there is a part of randomness, we generate ten training sets. The final models do not include the same variables as before. Not surprisingly, it depends on the 80% selected. The number of predictors is five or six. Nevertheless, whatever the training set, the meta path MT-UH is always the first predictor to be selected. After, there is no more consensus on the second variable but the RP-UH and RT-RP-UH always compete for the second place. Again, it is not surprising to obtain the RP-UH variable since, for a user, it is related to one of the closest neighbors with respect to our graph construction. Although the r^2 scores of the final models reach, on average, 0.7 for the training samples, we only get, on average, a score of 0.5 for the test sets (Fig. 5). One also observes that even if a model fits better the training set, it does not mean that it will give the best recovery. Indeed, it is sometimes better to consider a model with fewer regressors, and so a lower r^2 for training set, to better recover.

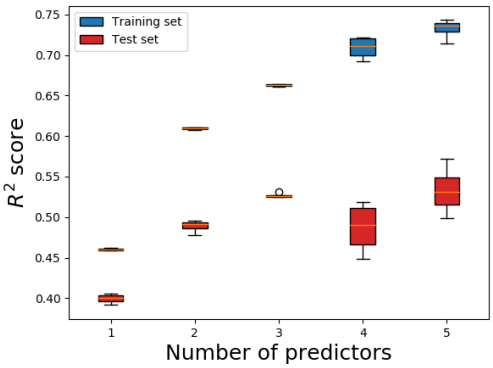

Figure 5: Boxplot for the r^2 scores of training sets and test sets. The training set scores increase with the number of predictors in the model while for the testing set, the scores seem to reach a threshold.

5. Related Work. As previously explained, our work is based on node similarity measures. Recently, several measures tackle the problem of node similarity in HIN which takes into account not only the structure similarity of two entities but also the metapaths connecting them. Amongst these measures, PathCount (PC,[9]) and Path Constrained Random Walk (PCRW,[6]) are the two most basic and gave birth to several extensions [1, 2, 5, 12].

The methods related to PC are based on the count of paths between a pair of nodes, given a meta path. PathSim [10] measures the similarity between two objects of same type along a symmetric meta path which is restrictive since many valuable paths are asymmetric and the relatedness between entities of different types is not useless. Two measures based on it [3, 4] incorporate more information such as degree nodes and transitivity. However, all these methods have the drawback of favoring highly connected objects since they deal with raw data.

The methods related to PCRW are based on random walks and so the probability of reaching a node from another one, given a meta path. Considering a random walk implies a normalization and, depending on the data, offers better results. An adaptation, HeteSim [8], measures the meeting probability between

two walkers starting from opposite extremities of a path, given a meta path. However, this method requires the decomposition of atomic relations, which is very costly for large graphs. To address this issue, AvgSim [11] computes the similarity between two nodes using random walks conditioned by a meta path and its inverse. But it is mostly appreciated in undirected networks since in these cases, it is just as sensible to walk a path in one direction as in the other.

In these cited works, when the similarity scores are used for link prediction/detection, the scores are ranked and then, the presence of links is inferred based on this ranking. Also some work try to combine meta paths but the target values to recover are binary; the networks are unweighted. At variance with these works, we set ourselves in the general framework of directed and weighted HINs. We do not use rankings but take directly the similarity measures obtained by means of an adequate combination of PCRWs as link weights. This allows not only to perform description tasks but also, to some extent, recovery tasks.

6. Conclusion and perspectives. We have considered a linear combination of path-constrained random walks to try to explain, to some extent, a specific meta path in a HIN. This proposed method allows to express the weight of a link between two nodes knowing the other links in a graph. This which could be useful for prediction or recommendation tasks. In particular, we have shown on our dataset, that the hashtags posted by a specific user is mainly related to those posted by her direct neighborhood, especially the MT and RP neighborhood. This method has also shown that the RT relation is not really useful for our purpose.

Nevertheless, the main drawback of the method is its sensitivity to outliers. Hence, more robust least square alternatives could be envisaged such that Least Trimmed Squares or parametric alternatives.

Furthermore, we have provided all the meta paths whose length is no longer than four. Even if it is motivated by previous tests, this threshold is clearly data related and is based on the knowledge of the user. Hence, it could be interesting to build a method able to find relevant meta paths by itself.

Finally, all data have been aggregated in time. Since it is possible to extract the time stamp of tweets, a future work could be the integration of time by defining a random walk process on temporal graph or by counting the temporal paths (plus normalization). This would restrict the possibilities of the walker and maybe improve the quality of the model.

Acknowledgement. This work is funded in part by the European Commission H2020 FETPROACT 2016-2017 program under grant 732942 (ODYCCEUS), by the ANR (French National Agency of Research) under grants ANR-15- E38-0001 (AlgoDiv), by the Île-de-France Region and its program FUI21 under grant 16010629 (iTRAC).

References

[1] Y. Fang, W. Lin, V. W. Zheng, M. Wu, K. C. Chang, and X. Li. Semantic proximity search on graphs with metagraph-based learning. In *2016 IEEE 32nd International Conference on Data Engineering (ICDE)*, pages 277–288, May 2016.

[2] M. Gupta, P. Kumar, and B. Bhasker. Dprel: A meta-path based relevance measure for mining heterogeneous networks. *Information Systems Frontiers*, Nov 2017.

[3] J. He, J. Bailey, and R. Zhang. Exploiting transitive similarity and temporal dynamics for similarity search in heterogeneous information networks. In *DASFAA*, 2014.

[4] L. Hou U., K. Yao, and H. F. Mak. Pathsimext: Revisiting pathsim in heterogeneous information networks. In F. Li, G. Li, S.-w. Hwang, B. Yao, and Z. Zhang, editors, *Web-Age Information Management*, pages 38–42, Cham, 2014. Springer International Publishing.

[5] Z. Huang, Y. Zheng, R. Cheng, Y. Sun, N. Mamoulis, and X. Li. Meta structure: Computing relevance in large heterogeneous information networks. In *Proceedings of the 22Nd ACM SIGKDD*

International Conference on Knowledge Discovery and Data Mining, KDD '16, pages 1595–1604, New York, NY, USA, 2016. ACM.

[6] N. Lao and W. W. Cohen. Relational retrieval using a combination of path-constrained random walks. *Mach. Learn.*, 81(1):53–67, October 2010.

[7] S.A. Macskassy. On the study of social interactions in twitter. In *Sixth International AAAI Conference on Weblogs and Social Media. ICWSM*, 2012.

[8] C. Shi, X. Kong, Y. Huang, P. S. Yu, and B. Wu. Hetesim: A general framework for relevance measure in heterogeneous networks. *IEEE Transactions on Knowledge & Data Engineering*, 26(10):2479–2492, Oct. 2014.

[9] Y. Sun, R. Barber, M. Gupta, C. C. Aggarwal, and J. Han. Co-author relationship prediction in heterogeneous bibliographic networks. In *Proceedings of the 2011 International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '11, pages 121–128, Washington, DC, USA, 2011. IEEE Computer Society.

[10] Y. Sun, J. Han, X. Yan, P. S. Yu, and T. Wu. Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. In *In VLDB' 11*, 2011.

[11] D. Xiao, X.Meng, Y. Li, C. Shi, and B. Wu. Avgsim: Relevance measurement on massive data in heterogeneous networks. 2016.

[12] Y. Zhou, J. Huang, H. Sun, and Y. Sun. Recurrent meta-structure for robust similarity measure in heterogeneous information networks. *ArXiv e-prints*, December 2017.