Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 38 (2014) 68 – 75

Procedia Computer Science

10th Italian Research Conference on Digital Libraries, IRCDL 2014

Assessing Document Relevance by modeling Citation Networks with Probabilistic Graphs

Teresa M.A. Basile[a,*], Nicola Di Mauro[a], Floriana Esposito[a]

[a]*Department of Computer Science, LACAM laboratory, University of Bari "Aldo Moro", Via Orabona,4, 70125 Bari, Italy*

Abstract

Paper citation networks are a traditional social medium for the exchange of ideas and knowledge. In this paper we use citation networks as a mean to assess both the importance of the citations of a paper and to identify relevant papers. We addressed these problems by modeling the citation network with a probabilistic graph useful to infer unknown links among the nodes representing papers. The proposed approach has been evaluated on three real world citation network whose results proved its validity.

Peer-review under responsibility of the Scientific Committee of IRCDL 2014
Keywords: Probabilistic graphs; citation networks.

1. Introduction

The study of networked data, including social networks, biological networks and information networks, is one of the major topic in the current scientific research in Computer Science. Recently there is a growing trend in the study of various types of *scholarly networks*, wherein a node usually denotes an academic entity, such as an article, a journal, or an author, and links usually denote relationships such as citation, co-authorship, co-citation, bibliographic coupling, or co-word. Google, for instance, recently introduced[1] the Scholar Updates tool, that analyzes the articles identified in a Scholar profile to propose new articles relevant to them by using a statistical model that incorporates also the citation graph between articles.

The need to have automatic tools for searching highly related publications in terms of research fields and topics is due to the difficulty of researchers to follow the rapid growth of a scientific research field. Literature search tools allow users to find relevant paper using key-word-based approaches that, however, return thousands or millions of relevant papers, making difficult for a researcher to concentrate on those publications closely related to his research.

The focus of this paper are the *citation networks*, networks of references among documents that can be modeled as a graph. Each node represents a paper of the network and there is a direct link from a paper x to a paper y whether the paper x cites the paper y. In particular, in this paper we use the citation networks as a mean to assess both the

* Corresponding author. Tel.: +39-0805442297; fax: +39-0805442297.
E-mail address: teresamaria.basile@uniba.it (Teresa M.A. Basile).

[1] http://googlescholar.blogspot.it/2012/08/scholar-updates-making-new-connections.html

Peer-review under responsibility of the Scientific Committee of IRCDL 2014
doi:10.1016/j.procs.2014.10.013

relevance of the citations of a paper and to identify similar papers. We addressed these problems by modeling the citation network with a *probabilistic graph* useful to infer unknown links.

Probabilistic graphs is an important research topic emerged in the last few years, connected with that of Statistical Relational Learning [1], that extends the graph structures with uncertainty [2,3,4,5,6]. With a probabilistic graph one can model structured domain, as with the classical graph structure, but with the advantage to also handle uncertain data. Uncertainty is modeled by means of *probabilistic edges* whose value quantifies the likelihood of the edge existence, or the strength of the link between the nodes it connects. Here the edges are not assumed to absolutely exist, but, adopting the *possible world semantics*, they may exist according to their own probability. Since we have to deal with probabilistic edges, then arises the problem of transposing the tasks in classical graph structure to this probabilistic setting.

As we will see later, the use of probabilistic graphs give us the possibility to solve the problem of *link prediction* [7], whose task may be formalized as follows. Given a networked structure (V, E) made up of a set of data instances V and a set of observed links E among some nodes in V, the task corresponds to predict how likely should exist an unobserved link between two nodes. In a bibliographic citation context unobserved link should correspond to possible citations or similarities among papers.

Two long-established different citation relations currently used to compute similarity between papers are the *bibliographic coupling* (BC), originally proposed in [8], and the *co-citation* (CC), proposed in [9]. The aim of BC is to identify, within a set of publications, groups that share a common intellectual background. The bibliographic coupling between two papers corresponds to the number of common citations, while a co-citation between two papers represents the number of articles that cite both. The former relationship is backward-based (i.e., it needs information lying in the past) and static (i.e., its value cannot change over time), while the latter is forward-based (i.e., two papers have to be cited by other authors to properly compute the similarity, and hence it could be inappropriate for recent papers).

Both BC and CC use only the information embedded in the reference list of each paper neglecting the networked information of a citation graph. For instance, two paper may be not related by a BC relationships but they could be strongly related if their local sub-networks have a lot in common. This consideration have just been explored in [10], where the authors present link-based similarity estimation methods on a citation graph based on connectivity alone to asses the relatedness between scientific papers.

The growing interest in predicting relevant papers using the citation graph is showed by the number of recent publications [10,11,12,13]. The key difference of this paper with respect to that of [10] is that we enrich the citation graph with bibliographic coupling relationships obtaining a more complex relational network that can be modeled as a probabilistic graph. In particular, we have certain directed edges connecting a paper x that cites a paper y and probabilistic symmetric edges between papers that are bibliographically coupled, whose probability denotes their relatedness. [11] proposes a supervised machine learning approach, based on some papers information such as authors, topics, target publication venues and publication time, to solve the citation prediction problem (i.e., predicting the citation relationship between a query paper and a set of previous papers). [12] proposes a supervised approach to classify the citation relation between papers. Both [11] and [12] need context information that in the general case is not available. Our proposed approach only uses the structural information of the citation graph without exploiting other natural language processing technique to assess similarities among the papers.

The paper is organized as follows. Section 2 introduces the basics of probabilistic graphs as in [5,6]. Section 3 presents the validation of the proposed method on three real world citation graph. Finally, Section 4 concludes the paper.

2. Probabilistic Graphs

Let $G = (V, E)$, be a graph where V is a collection of nodes and $E \subseteq V \times V$ is the set of edges, or relationships, between the nodes.

Definition 2.1 (Probabilistic graph)**.** A *probabilistic graph* is a system $G = (V, E, \Sigma, l_V, l_E, s, t, p)$, where (V, E) is an directed graph, V is the set of nodes, E is the set of ordered pairs of nodes where e=(s,t), Σ is a set of labels, $l_V : V \to \Sigma$ is a function assigning labels to nodes, $l_E : E \to \Sigma$ is a function assigning labels to the edges, $s : E \to V$ is a function returning the source node of an edge, $t : E \to V$ is a function returning the target node of an edge, $p : E \to [0, 1]$ is a function assigning *existence probability* values to the edges.

The existence probability $p(a)$ of an edge $a = (u, v) \in E$ is the probability that the edge a, connecting the node u to the node v, can exist in the graph. A particular case of probabilistic graph is the *discrete graph*[2], where binary edges between nodes represent the presence or absence of a relationship between them, i.e., the existence probability value on all observed edges is 1.

The *possible world semantics*, specifying a probability distribution on discrete graphs and formalized in the *distribution semantics* of Sato[14] for the first order logic, is usually used for probabilistic graphs. We can imagine a probabilistic graph G as a sampler of worlds, where each world is an instance of G. A discrete graph G' is sampled from G according to the probability distribution P, denoted as $G' \sqsubseteq G$, when each edge $a \in E$ is selected to be an edge of G' with probability $p(a)$. Edges labeled with probabilities are treated as mutually independent random variables indicating whether or not the corresponding edge belongs to a discrete graph.

Assuming independence among edges, the probability distribution over discrete graphs $G' = (V, E') \sqsubseteq G = (V, E)$ is given by

$$P(G'|G) = \prod_{a \in E'} p(a) \prod_{a \in E \setminus E'} (1 - p(a)). \tag{1}$$

Definition 2.2 (Simple path)**.** Given an uncertain graph G, a *simple path* of length k from u to v in G is a sequence of edges denoted as $\pi = \langle e_1, e_2, \dots e_k \rangle$, where $e_1 = (u, v_1)$, $e_k = (v_{k-1}, v)$, and $e_i = (v_{i-1}, v_i)$ for $1 < i < k - 1$, or equivalently as $\pi = u \xrightarrow{e_1} v_1 \xrightarrow{e_2} v_2 \cdots v_{k-1} \xrightarrow{e_k} v$.

Given an uncertain graph G, and $\pi = \langle e_1, e_2, \dots e_k \rangle$ a path in G from the node u to the node v, $\ell(\pi) = l_E(e_1)l(e_2) \cdots l(e_k)$ denotes the ordered concatenation of the labels of all the edges belonging π. As in[6], we adopt a *regular expression* $\mathtt{R}$ to denote what is the exact sequence of the labels that the path must contain.

Definition 2.3 (Language-constrained simple path)**.** Given a probabilistic graph G and a *regular expression* $\mathtt{R}$, a *language constrained simple path* is a simple path π such that $\ell(\pi) \in L(\mathtt{R})$, where $L(\mathtt{R})$ is the language described by $\mathtt{R}$.

2.1. Inference

Given a probabilistic graph G, a main task corresponds to compute the probability that there exists a simple path between two nodes u and v, that is, querying for the probability that a randomly sampled discrete graph contains a simple path between u and v. More formally, the *existence probability* $P(\pi|G)$ of a simple path π in a probabilistic graph G corresponds to the marginal $P((\pi, G')|G)$ with respect to π:

$$P(\pi|G) = \sum_{G' \sqsubseteq G} \mathbb{1}\{\pi \in G'\} \cdot P(G'|G), \tag{2}$$

where $\mathbb{1}\{\pi \in G'\} = 1$ if there exits the simple path π in G', and $\mathbb{1}\{\pi \in G'\} = 0$ otherwise. In other words, the existence probability of the simple path π is the probability that the simple path π exists in a randomly sampled discrete graph.

Definition 2.4 (Language-constrained simple path probability)**.** Given a probabilistic graph G and a *regular expression* $\mathtt{R}$, the probability of a *language-constrained simple path* π is

$$P(\pi|G, \mathtt{R}) = \sum_{G' \sqsubseteq G} \mathbb{1}\{\pi \in G'|\mathtt{R}\} \cdot P(G'|G), \tag{3}$$

where $\mathbb{1}\{\pi \in G'|\mathtt{R}\} = 1$ if there exists a simple path π in G' such that $\ell(\pi) \in L(\mathtt{R})$, and $\mathbb{1}\{\pi \in G'|\mathtt{R}\} = 0$ otherwise.

The previous definition give us the possibility to compute the probability of a set of simple path queries, or patterns, fulfilling the structure imposed by a regular expression, as in[6]. In this way we are interested in discrete graphs that contain at least one simple path belonging to the language denoted by the regular expression.

Computing the existence probability directly using (2) or (3) is intensive and intractable for large graphs since the number of discrete graphs to be checked is exponential in the number of probabilistic edges. It involves computing the existence of the simple path in every discrete graph and accumulating their probability.

[2] Sometimes called *certain graph*.

	HEP-TH	HEP-PH	PATENTS
Nodes	27,770	34,546	3,774,768
Edges	352,807	421,578	16,518,948

Table 1. Statistics of the HEP-PH, HEP-TH and PATENTS datasets.

A natural way to overcome the intractability of computing the existence probability of a simple path is to approximate it using a Monte Carlo sampling approach [15]:

1. we sample n possible discrete graphs, $G_1, G_2, \ldots G_n$ from G by sampling edges uniformly at random according to their edge probabilities; and
2. we check if the simple path exists in each sampled graph G_i.

This process provides the following basic sampling estimator for $P(\pi|G)$:

$$P(\pi|G) \approx \widehat{P(\pi|G)} = \frac{\sum_{i=1}^{n} \mathbb{1}\{\pi \in G_i\}}{n}. \tag{4}$$

Note that is not necessary to sample all the edges to check whether the graph contains the path. For instance, assuming to use an iterative depth first search (DFS) procedure to check the path existence. When a node is just visited, we will sample all its adjacent edges and pushing them into the stack used by the iterative procedure. We will stop the procedure either when the target node is reached or when the stack is empty (non existence).

Algorithm 1 reports the algorithm to solve the inference step. The function `sampledAsTrue` implements a memoization technique in order to sample the edges. If the function is called for the first time on a given edge e, then it sample the edge and returns true whether the edge has been sampled as true and false otherwise. All successive calls on the same already sampled edge consist in returning the previous sampled value. The algorithm corresponds to a DFS starting from the node u and ending to the node v if possible. If the search ends in v a positive count is accumulated. Then the estimated probability is computed by dividing the accumulated positive count by the number of samplings n.

Given a probabilistic graph G with d the mean degree of its node, and k the length of a path π, then the complexity of the Algorithm 1 is $O(pd^k)$, where p is the mean probability of the edges.

3. Experimental evaluation

In order to evaluate our proposed approach we used three publicly available datasets [16,17]: the arXiv High Energy Physics Teory (HEP-TH) dataset[3], the Arxiv High Energy Physics Phenomenology (HEP-PH) dataset[4], both originally released as part of KDD Cup 2003 competition[5], and the U.S. patent (PATENTS) dataset[6]. Table 1 reports the statistics of the three citation graphs we used.

HEP-TH (resp., HEP-PH) is a citation graph from the e-print arXiv, covering all the citations within a dataset of 27,770 (resp., 34,546) papers with 352,807 (resp., 421,578) edges. If a paper i cites a paper j, the graph contains a directed edge from i to j. If a paper cites, or is cited by, a paper outside the dataset, the graph does not contain any information about this. The data for both HEP-TH and HEP-PH covers papers in the period from January 1993 to April 2003 (124 months). It begins within a few months of the inception of the arXiv, and thus represents essentially the complete history of its HEP-TH and HEP-PH sections.

The U.S. patent dataset is maintained by the National Bureau of Economic Research[7]. The data set spans 37 years (January 1, 1963 to December 30, 1999), and includes all the utility patents granted during that period, totaling

[3] http://snap.stanford.edu/data/cit-HepTh.html
[4] http://snap.stanford.edu/data/cit-HepPh.html
[5] http://www.cs.cornell.edu/projects/kddcup/index.html
[6] http://snap.stanford.edu/data/cit-Patents.html
[7] http://www.nber.org/patents/

Algorithm 1 INFER(G, π, R, n)

Input: G: the probabilistic graph; π: the path $u \xrightarrow{e_1} v_1 \xrightarrow{e_2} v_2 \cdots v_{k-1} \xrightarrow{e_k} v$; R: the regular expression; n: the number of samplings;

Output: $\widehat{P(\pi|G, \mathrm{R})}$

```
1:  c = 0
2:  for i = 1 to n do
3:      visited = { u }
4:      S.clear()
5:      sampler.clear()
6:      depth = 1
7:      prevNode[0] = u
8:      proven = false
9:      for all adjacent node a_j of the node u do
10:         S.push((a_j,depth))
11:     while not S.empty() and not proven do
12:         (a, depth) = S.top()
13:         S.pop()
14:         e = (prevNode[depth-1],a)
15:         prevNode[depth] = a
16:         if a ∉ visited and ℓ(e) == e_depth then
17:             sampled = sampledAsTrue(e, sampler)
18:             if depth == k then
19:                 c++
20:                 proven = true
21:             visited.add(a)
22:             for all adjacent node a'_j of the node a do
23:                 S.push((a'_j,depth+1))
24: return c/n
```

3, 923, 922 patents. The citation graph includes all citations made by patents granted between 1975 and 1999, totaling 16, 522, 438 citations.

3.1. Probabilistic graph construction

The considered HEP-TH, HEP-PH and PATENTS citation graph does not contain probabilistic edges, the probability of the citation edges here are set to 1. In order to apply the proposed method, we enriched the original graph by adding bibliographic coupling relationships, with their probabilistic value, among the papers in the following way. For each pair of papers a and b of the dataset we computed their coupling strength using the Pianka's index [18]:

$$p_{ab} = \frac{C_a \cap C_b}{(C_a C_b)^{1/2}}, \tag{5}$$

where C_a and C_b are, respectively, the citations of the paper a and b. If the coupling strength is greater than zero we add a probabilistic edge between a and b with probability equal to p_{ab}.

Hence the citation graph contains node denoting specific papers ($\mathtt{paper}_i$), the citation links connecting two papers ($\mathtt{paper}_i \xrightarrow{\mathtt{cite}} \mathtt{paper}_j$), and the mutual links connecting two paper with a positive coupling strength ($\mathtt{paper}_i \xleftrightarrow{\mathtt{coupling}} \mathtt{paper}_j$).

Once have constructed the probabilistic graph G, in order to predict the existence probability of a link between two papers the following regular expressions have been used:

$$\mathrm{R}_1 = \{\mathtt{cite}^1, \mathtt{coupling}^1\}$$

Dataset	papers	citations	nodes	edges	cite	coupling
HEP-TH	76	17.0	2,005.4	12,310.3	24.1%	75.9%
HEP-PH	35	16.9	3,439.9	22,654.9	32.3%	67.7%
PATENTS	100	16.5	1,787.6	504,193.0	44.8%	55.2%

Table 2. Statistics of the selected paper from the HEP-TH, HEP-PH and PATENTS datasets.

and

$$\mathtt{R}_2 = \{\mathtt{coupling}^1, \mathtt{cite}^1\},$$

corresponding, respectively to use the simple path $\pi_1 = \mathtt{paper}_i \xrightarrow{\mathtt{cite}} \mathtt{paper}_j \xrightarrow{\mathtt{coupling}} \mathtt{paper}_k$ and $\pi_2 = \mathtt{paper}_i \xrightarrow{\mathtt{coupling}} \mathtt{paper}_j \xrightarrow{\mathtt{cite}} \mathtt{paper}_k$. The link, in this case indicating that the two paper are similar, is predicted to exist with a given probability by solving the inference step $P(\pi_i|\mathtt{R}_i, G)$ over the probabilistic graph. In particular, given two paper a and b, if the probability $P(\pi_1|\mathtt{R}_1, G) = P(\mathtt{a} \xrightarrow{\mathtt{cite}} \mathtt{paper}_j \xrightarrow{\mathtt{coupling}} \mathtt{b}|\mathtt{R}_1, G)$ is very high we could assume that the paper b is highly relevant for the paper a, and *viceversa*.

The first path π_1 is useful to assess only the relevance of the cited papers, since, given a starting paper, the first step is to reach one of its cited papers. In particular, given a paper a, the relevance of a cited paper b is computed by seeing how the other cited papers are bibliographically coupled with respect to b.

By using the second path π_2, we are able to compute the relevance of both cited and other non cited papers. Given a paper a, the relevance of a non cited paper b is computed by seeing whether the other bibliographically coupled papers cite b. More complex paths may be used to inspect more in depth the neighborhood of a paper.

3.2. Results

Among all the papers contained in the datasets HEP-TH and HEP-PH we chosen to validate the approach only those of the year 2002 having an number of citation belonging to the interval [15,19] and with a corresponding sub-network (the considered paper, its cited papers, the bibliographically coupled papers and their citations) having a number of nodes lesser than 5000, thus leading to consider 76 papers for HEP-TH and 35 papers for HEP-PH. While for the patents dataset we randomly chosen 100 patents having an number of citation belonging to the interval [15,19] and with a corresponding sub-network having a number of nodes lesser than 5000.

Table 2 reports some statistics about the considered papers whose citations are 17, 16.9 and 16.5 on average, respectively for the HEP-TH, HEP-PH and PATENTS dataset. Each local subnetwork regarding a paper contains 2, 005.4 (resp., 3, 439.9 and 1, 787.6) nodes and 12, 310.3 (resp., 22, 654.9 and 504, 193) edges on average for the HEP-TH (resp., HEP-PH and PATENTS) dataset; among all the edges 24.1% (resp., 32.3% and 44.8%) are citation links and 75.9% (resp., 67.7% and 55.2%) corresponds to bibliographic coupling relationships for the HEP-TH (resp., HEP-PH and PATENTS) dataset.

Table 3 reports the citations of the paper titled "Null string evolution in black hole and cosmological spacetimes" by M.P. Dabrowski and I.Prochnicka, ranked by using the two different paths π_1 and π_2, where the authors discuss the problem of the motion of classical strings in some black hole and cosmological spacetimes. Rankings obtained by following the paths of the kind π_2 seem to be better that those obtained with the path π_1, i.e., checking the relevance of a paper by looking how bibliographically coupled papers cite it is more effective than verifying how other cited papers are bibliographically coupled with it.

In order to evaluate the accuracy of the proposed approach we tested how the results returned making inference over the probabilistic citation graph are relevant with respect to a given paper.

Given C_p the set of bibliographically coupled papers of a given paper p, in this experiment we used the path π_2 to find: a) relevant papers to p but not bibliographically coupled with it (π_2^a), and b) relevant papers to p that are bibliographically coupled with it (π_2^b). For instance in the HEP-TH dataset, each paper is on average bibliographically coupled with 99.7 other papers, and from these papers we can reach on average, by citation, 840.1 other papers. Hence, starting from the node p, using the path π_2^a (resp., π_2^b) we reach one of its bibliographically coupled paper that does not cite (resp., cites) a paper that is bibliographically coupled with p. All the possible nodes that can be reached with this path (939.8 on average) are tested for relevance and ranked according to their probability.

π_2	π_1	Title
1.0	0.0	**Strings** in **Cosmological** and **Black Hole** Backgrounds: Ring Solutions (1993)
1.0	0.365	Circular **String**-Instabilities in Curved **Spacetime** (1993)
1.0	0.884	**Strings** Propagating in the 2+1 Dimensional **Black Hole** Anti de Sitter **Spacetime** (1994)
1.0	0.982	New Classes of Exact Multi-**String** Solutions in Curved **Spacetimes** (1995)
1.0	0.0	Friedmann Universes and Exact Solutions in **String Cosmology** (1995)
0.999	0.994	Small E_8 Instantons and Tensionless Non-critical **Strings** (1996)
0.997	0.34	Schild's **Null Strings** in Flat and Curved Backgrounds (1995)
0.997	0.91	Tension as a Perturbative Parameter in Non–Linear **String** Equations in Curved **Space-Time** (1996)
0.996	0.966	Planetoid **String** Solutions in 3 + 1 Axisymmetric **Spacetimes** (1996)
0.989	0.0	AdS Dynamics from Conformal Field Theory (1998)
0.952	0.971	**Strings** in Homogeneous Background **Spacetimes** (1997)
0.914	0.968	Variational principle and a perturbative solution of non-linear **string** equations in curved space (1998)
0.899	0.999	**Null Strings** in Kerr **Spacetime** (1997)
0.719	0.968	The Effect of Spatial Curvature on the Classical and Quantum **Strings** (1995)
0.423	0.973	Exact **String** Solutions in Nontrivial Backgrounds (2001)
0.186	0.998	Perturbative **String** Dynamics Near the Photon Sphere (1998)

Table 3. Citations ranking of the paper "Null string evolution in black hole and cosmological spacetimes (2002)" obtained by using the two different paths π_1 and π_2.

Dataset	Method	10	20	30	40	50	100
HEP-TH	coup.	0.240	0.384	0.467	0.529	0.578	0.681
	π_2^a	0.476	**0.725**	0.815	0.843	0.867	0.916
	π_2^b	**0.492**	0.709	**0.817**	**0.867**	**0.899**	**0.947**
HEP-PH	coup.	0.126	0.224	0.290	0.355	0.407	0.543
	π_2^a	**0.513**	**0.739**	**0.826**	**0.861**	**0.887**	0.934
	π_2^b	0.511	0.715	0.786	0.848	0.867	**0.941**
PATENTS	coupling	0.047	0.084	0.105	0.125	0.143	0.195
	π_2^a	0.485	0.697	0.752	0.782	0.815	0.865
	π_2^b	**0.542**	**0.816**	**0.883**	**0.911**	**0.922**	**0.928**

Table 4. Ranking results on the HEP-TH, HEP-PH and PATENTS datasets of a simple bibliographic coupling-based method and our proposed approach.

Table 4 reports the results of this experiment. The first row denotes the number of the first top k ranked papers for each method we adopted, ranging in the set $\{10, 20, 30, 40, 50, 100\}$. The baseline method (reported in the second row) is a simple bibliographic coupling-based approach that builds the ranked list by using the bibliographic strength as a score. The values in the table denote the percentage, on average over the 76 (resp., 35 and 100) papers for the HEP-TH (resp., HEP-PH and PATENTS) dataset, of the top k ranked papers that are cited by the considered papers. As we can see, the method that uses the path π_2^b always outperforms the baseline approach, as the path π_2^a, thus proving the validity of the proposed approach.

4. Conclusion

Paper citation graphs are a traditional social medium for the exchange of ideas and knowledge among researchers. In this paper we used citation graphs as a mean to assess both the importance of the citations included in a paper and to identify relevant papers. These problems have been addressed by modeling the citation graph with a probabilistic graph useful to infer unknown links among the nodes. The proposed approach has been evaluated on a real world citation network whose results proved its validity.

Acknowledgements

This work has been partially founded by the PON02 00563 3489339 project PUGLIA@SERVICE - L'Ingegneria dei Servizi Internet-Based per lo sviluppo strutturale di un territorio intelligente funded by the Italian Ministry of University and Research (MIUR). Furthermore, we would like to acknowledge the support of the European Commission through the project MAESTRA - Learning from Massive, Incompletely annotated, and Structured Data (Grant number ICT-2013-612944).

References

1. Getoor, L., Taskar, B.. *Introduction to Statistical Relational Learning (Adaptive Computation and Machine Learning)*. The MIT Press; 2007. ISBN 0262072882.
2. Potamias, M., Bonchi, F., Gionis, A., Kollios, G.. k-nearest neighbors in uncertain graphs. *Proc VLDB Endow* 2010;**3**:997–1008.
3. Zou, Z., Gao, H., Li, J.. Discovering frequent subgraphs over uncertain graph databases under probabilistic semantics. In: *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM; 2010, p. 633–642.
4. Pfeiffer III, J.J., Neville, J.. Methods to determine node centrality and clustering in graphs with uncertain structure. In: *Proceedings of the Fifth International Conference on Weblogs and Social Media*. The AAAI Press; 2011, .
5. Taranto, C., Di Mauro, N., Esposito, F.. Learning in probabilistic graphs exploiting language-constrained patterns. In: Appice, A., Ceci, M., Loglisci, C., Manco, G., Masciari, E., Ras, Z., editors. *New Frontiers in Mining Complex Patterns*; vol. 7765 of *LNCS*. Springer Berlin Heidelberg; 2013, p. 155–169.
6. Di Mauro, N., Taranto, C., Esposito, F.. Link classification with probabilistic graphs. *Journal of Intelligent Information Systems, DOI 101007/s10844-013-0293-0* 2014;doi:10.1007/s10844-013-0293-0.
7. Getoor, L., Diehl, C.P.. Link mining: a survey. *SIGKDD Explorations* 2005;**7**(2):3–12.
8. Kessler, M.. Bibliographic coupling between scientific papers. *American Documentation* 1963;**14**:10–25.
9. Small, H., Griffith, B.C.. The structure of scientific literatures i: Identifying and graphing specialties. *Science Studies* 1974;**4**(1):17–40.
10. Lu, W., Janssen, J., Milios, E., Japkowicz, N., Zhang, Y.. Node similarity in the citation graph. *Knowl Inf Syst* 2006;**11**(1):105–129.
11. Yu, X., Gu, Q., Zhou, M., Han, J.. Citation prediction in heterogeneous bibliographic networks. In: *SDM*. SIAM / Omnipress; 2012, p. 1119–1130.
12. Liang, Y., Li, Q., Qian, T.. Finding relevant papers based on citation relations. In: Wang, H., Li, S., Oyama, S., Hu, X., Qian, T., editors. *Web-Age Information Management*; vol. 6897 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg. ISBN 978-3-642-23534-4; 2011, p. 403–414.
13. Beel, J., Langer, S., Genzmehr, M., Gipp, B., Breitinger, C., Nrnberger, A.. Research paper recommender system evaluation: A quantitative literature survey. 2013.
14. Sato, T.. A statistical learning method for logic programs with distribution semantics. In: *In Ppoceedings of the 12th International Conference on Logic Programming (ICLP95*. MIT Press; 1995, p. 715–729.
15. Jin, R., Liu, L., Ding, B., Wang, H.. Distance-constraint reachability computation in uncertain graphs. *Proc VLDB Endow* 2011;**4**:551–562.
16. Gehrke, J., Ginsparg, P., Kleinberg, J.M.. Overview of the 2003 kdd cup. *SIGKDD Explorations* 2003;**5**(2):149–151.
17. Leskovec, J., Kleinberg, J., Faloutsos, C.. Graphs over time: densification laws, shrinking diameters and possible explanations. In: *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM; 2005, p. 177–187.
18. Pianka, E.. The structure of lizard communities. *Ann Rev Ecol Syst* 1973;**4**:53–74.