Supervised Machine Learning Applied to Link Prediction in Bipartite Social Networks

Conference Paper · August 2010

DOI: 10.1109/ASONAM.2010.87 · Source: DBLP

3 authors, including:

Rushed Kanawati
Université Paris 13 Nord
143 PUBLICATIONS **764** CITATIONS
SEE PROFILE

Céline Rouveirol
Université Paris 13 Nord
93 PUBLICATIONS **1,304** CITATIONS
SEE PROFILE

2010 International Conference on Advances in Social Networks Analysis and Mining

Supervised Machine Learning applied to Link Prediction in Bipartite Social Networks

Nesserine Benchettara, Rushed Kanawati, Céline Rouveirol
LIPN - CNRS UMR 7030
University of Paris Nord
99 Av. J.B. Clément
93430 Villetaneuse, FRANCE
Email:firstname.lastname@lipn.univ-paris13.fr

***Abstract*—This work copes with the problem of link prediction in large-scale two-mode social networks. Two variations of the link prediction tasks are studied: predicting links in a bipartite graph and predicting links in a unimodal graph obtained by the projection of a bipartite graph over one of its node sets. For both tasks, we show in an empirical way, that taking into account the bipartite nature of the graph can enhance substantially the performances of prediction models we learn. This is achieved by introducing new variations of topological atttributes to measure the likelihood of two nodes to be connected. Our approach, for both tasks, consists in expressing the link prediction problem as a two class discrimination problem. Classical supervised machine learning approaches can then be applied in order to learn prediction models. Experimental validation of the proposed approach is carried out on two real data sets: a co-authoring network extracted from the DBLP bibliographical database and bipartite graph 8-years history of transactions on an on-line music e-commerce site.**

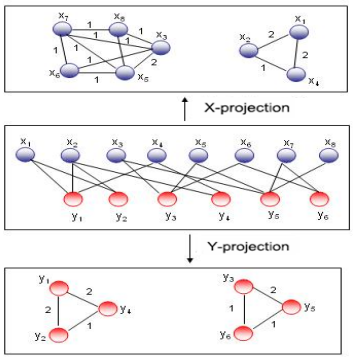

Fig. 1. Projections of a bipartite graph

I. Introduction

Many of naturally arising networks can either be directly modeled as bipartite graphs such as *affiliation networks*, or are obtained by the projection of a bipartite graph such as *co-membership* networks. A bipartite graph G is defined as follows: $G =< X, Y, E >$ where X and Y are two mutually exclusive sets of nodes. E is a set of edges of G and is a subset of $X \times Y$. A unimodal graph can be obtained from a bipartite graph by projecting the graph over one of its nodes' sets. For example, the projection over the X set is defined by a unimodal graph where nodes from X are tied if they are linked to at least n common nodes in the initial bipartite graph G. In a more formal way, let $\Gamma_g(x)$ be the set of neighbors of node x in a graph g. Projections of a bipartite graph G are then defined as follows (cf. figure 1):

- $G_X^n =< V_X \subseteq X, E = \{(x,y) : x,y \in X, |\Gamma_G(x) \cap \Gamma_G(y)| \geq n\} >$
- $G_Y^m =< V_Y \subseteq Y, E = \{(x,y) : x,y \in Y, |\Gamma_G(x) \cap \Gamma_G(y)| \geq m\} >$

In this work, we report on our experience in developing a new approach for link prediction in two types of graphs:

- bipartite graphs,
- unimodal graphs obtained by a projection of a bipartite graph.

The link prediction problem can be defined as follows [1]. Let $< G_1, G_2, \ldots, G_t >$ be a temporal sequence of networks, the link prediction problem consists in predicting what new links will appear at time $t+1$ linking nodes that have never been linked before. Our work is motivated by two applications:

- *Product recommendation in an e-commerce site*[1]. Actually, purchase data can be represented by a bipartite graph linking customers to items they've purchased. Product recommendation problem is naturally expressed as link prediction problem in this bipartite graph. The need for specific atributes for analyzing bipartite graphs has beed pointed out in several earlier works [2], [3]. A link-prediction approach in two-mode networks for product recommendation is also proposed in [4]. In contrast to previously existing approaches, we propose to compute the likelihood of a link in a bipartite graph applying metrics computed in both projected graphs (with different values of the projection parameters n and m).
- *Academic collaboration recommendation*[2]. The goal here is to recommend new academic collaborations by mining publicly available bibliographical databases. We have worked here on exploring *co-authorship* networks. Co-authorship networks have been extensively studied in the

[1]This application is part of the CADI project, with a grant from the French National Agency of Research (ANR) 2008-2010
[2]This application is part of the LIPS project with a BQR grant from the University Paris Nord in 2009

978-0-7695-4138-9/10 $26.00 © 2010 IEEE
DOI 10.1109/ASONAM.2010.87

field of social network analysis [5], [6]. Some earlier link prediction approaches have also been applied to this type of data [7], [8], [9]. However, none of these earlier works take into account that a co-authorship network is actually a projected graph (i.e., the projection of a bipartite graph linking authors to articles). We propose here new metrics for computing the likelihood of a link to be computed in the co-authorship network but also in the dual graph obtained by projecting the original bipartite graph over the publication set.

The task of recommending products to a customer fits naturally in this bipartite graph framework. In this case, X is the set of customers, Y is the set of items, G_X^n is the graph connecting customers that have bought at least n items in common, G_Y^m is the graph connecting items that have been bought by at least m common customers. Recommending products consists in predicting the occurrence of a new link in the initial bipartite graph. Recommending an academic collaboration is less intuitively expressed in a bipartite graph context. In order to recommend a new academic collaboration, one has to predict the occurrence of a new link in a projected graph, namely the co-author graph. Again, we propose to explore the initial bipartite graph as well as the projected publication graph for that purpose.

Different link prediction approaches have been proposed in the scientific literature. A very quick overview of existing methods is presented in section II. The approach we have applied follows the same general idea of works presented in [8], [9], [10], where the link prediction problem is expressed as two-class discrimination problem allowing hence to apply classical supervised machine learning approach for learning prediction models. For the two above recommendation task, we show through experimentations on real bibliographical and e-commerce datasets that taking into account the original bipartite nature of the publication network can substantially enhance the performances of obtained link prediction models. Experiments and obtained results are described and commented in section IV. Finally, we conclude in section V.

II. Link prediction approaches

We propose to classify the different approaches for link prediction in social networks according to the following three criteria:

- *Dyadic vs. Structural approaches*. A dyadic approach consists in computing a link *score* for each couple of unlinked vertices in graphs G_i (target vertices should be in the same connected component). The majority of link prediction approaches apply a dyadic approach. Different systems apply different score functions: topological scores [7], node features based scores [11], [12] or hybrid functions combining both types [13]. In contrast, structural approaches mine rules of evolution of (frequent) sub-graphs in the whole network. Frequent subgraph mining approaches can be used for that purpose [14], [15]. Such approaches can predict the simultaneous appearance (as well as disappearance) of multiple links.
- *Topological vs. node features based approaches*. Score function of a link can be based on mere topological features such the degree, the common neighbors, the distance between involved nodes, etc. or can be dependent of the node features itself. Topological approaches present the advantage of being application-independent.
- *Temporal vs. non temporal approaches*. Most approaches ignore the temporal aspects of network formation. In other words, a link established at time $t-i$ is considered having the same effect, for the prediction task, as a link established at time t. Recently, several works proposed to take the temporal issue into consideration. Examples are works presented in [16], [17].

In this work, we apply a dyadic topological approach. This will be sketched in the next section.

III. Proposed approach

Similarly to works proposed in [8], [9], [10], we formulate the link prediction as a supervised learning problem. The goal is to discriminate between examples of the `linked` class (positive examples) against examples of the `not-linked` class (negative examples). Learning such a supervised classification model requires building a training data that describes *examples* of both classes. For the sake of clarity, we explain hereafter how to generate training data for the link prediction problem in the case of a simple unimodal graph. Then, we show how to adapt this approach to graphs that have an underlying bipartite structure.

A. Supervised machine learning approach

The training data is constructed as follows. Let G_{obs} be a graph that summarizes in some way the temporal sequence of networks $G =< G_1, \ldots, G_t >$, and let us refer to G_{t+1}, as the *labelling* graph. In this work, as in many pther ones, G_{obs} is computed as the union of all snapshots in the sequence G. An example will be generated for each couple of nodes (x, y) such that:

- x and y belong to both G_{obs} and G_{t+1}
- $< x, y >$ is not an edge of G_{obs}
- x and y belong to the same connected component of G_{obs}

For each couple of nodes (x, y) satisfying the previous conditions, an example is built, containing the following information:

- a set of topological attributes computed for (x, y). In section III-B, we detail a list of topological attributes described in the literature,
- the class label associated to the example: if the couple of nodes (x, y) are actually linked in the labeling graph G_{t+1}, (x, y) is a positive example, it is a negative example otherwise.

A supervised learning technique is then applied to compute a predictive model that discriminates positive examples from negative ones. The whole process for constructing a supervised learning dataset is summarized figure 2.

As classical in a supervised learning context, the predictive models are then evaluated on examples which have not been seen by the system while learning. One of the most used way to evaluate predictive models in machine learning is cross

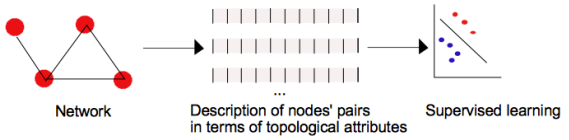

Fig. 2. Reformulating link prediction as a supervised learning task.

validation. One round of cross-validation involves partitioning a sample of data into two complementary subsets, learning a predictive model on one subset of the initial dataset (the training set), and then evaluating the obtained model on the complement subset (the validation set). To reduce variability when estimating predictive performance, multiple rounds of cross-validation (10 fold cross-validation, most often) are performed using different folds of a given partition, and averaging the validation results over the folds. The evaluation measures used in this work are:

- Precision = $\frac{\#\text{True positive examples}}{\#\text{Predicted positive examples}}$, where True Positive examples are examples predicted as positive and which are Real Positives (i.e., linked in the labelling graph)
- Recall = $\frac{\#\text{True positive examples}}{\#\text{Real positive examples}}$
- F1-measure = $2 \times \frac{Precision \times Recall}{Precision + Recall}$

We show in section IV that the supervised learning problems extracted from both recommendation applications are difficult problems. Actually, obtained example sets are very skewed: the number of positive examples is very small compared to negative ones (since few links do actually appear in the labeling graph). The difficulty is to identify the positive class, and as a consequence, we only provide in our evaluation the predictive performance of our classifier for the positive class.

B. Topological attributes

For each such couple of nodes, we compute a set of topological attributes that characterize their roles in the network as well as their *similarity*. We identify three types of topological attributes that can be computed. These will be described briefly in next subsections. First, we introduce some basic notations that help in describing the different topological attributes. Let $\Gamma_G(x)$ be the set of direct neighbors of a node x in a graph G. The set of neighbors of a node x is denoted $\Gamma(x)$ when there is no ambiguity concerning the considered graph. $|E|$ refers to the cardinality of the set E. Graphs handled in our study are not oriented. Hence, the degree of a node x in a graph G is equal to $|\Gamma_G(x)|$. Metrics describing pairs of nodes in a graph can be divided into three groups, detailed in the following.

1) Product of node's topological attributes: Examples of single node topological features (or attributes) are: the degree, the PageRank [18], the *hub* and the *authority* attributes computed by the HITS algorithm [19], the relative position of the node in the connected component: a central or a peripheral node, etc. A link score can be estimated by the product of such attributes for the two involved nodes. A well known example is the preferential attachment model [20]. Other known examples of node's topological features are the different centrality measures [21].

2) Neighborhood based metrics: The most frequently used neighborhood based attributes in the literature are the following:

1) Common Neighbours (denoted $CN(x,y)$): $|\Gamma(x) \cap \Gamma(y)|$,
2) Jaccard's coefficient (denoted $JC(x,y)$): $\frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|}$,
3) Preferential Attachment (denoted $PA(x,y)$): $|\Gamma(x) \times \Gamma(y)|$ [4],
4) a measure proposed in [22] (denoted $AA(x,y)$): $\Sigma_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{log(|\Gamma(z)|)}$.

3) Distance-based metrics: The simplest attribute to compute is the shortest path between x and y, denoted by $SP(x,y)$. Another such attribute, also quite frequently used in affiliation analyses between nodes of a graph, is the measure proposed in [23]: it consists in computing a weighted sum of all paths between x and y. More formally, the score of a link $< x, y >$ is computed as follows: $Katz(x,y) = \Sigma_{l=1}^{\infty} \beta^{\ell} \times |path_{x,y}^{(\ell)}|$, where $path_{x,y}^{(\ell)}$ is the number of paths between x and y of length ℓ. β is a positive parameter which favors shortest paths.

Other measures based on random walks through the considered graph [24] or exploiting the analogy with electrical circuits [25] are also proposed in the literature.

C. Link prediction in a bipartite graph

In the case of predicting links in a bipartite graph, for each simple topological attribute computed in the bipartite graph, two indirect attributes are computed in each projected graph. Let $a_G^d(x \in X, y \in Y)$ be a direct topological attribute computed in G. For each such attribute, we introduce the following two indirect attributes: $\varphi_{u \in \Gamma_G(y)} a^i_{G_X^n}(x,u)$ and $\varphi_{v \in \Gamma_G(x)} a^i_{G_Y^m}(v,y)$. Indirect attributes built from a are computed in both projected graphs G_X^n and G_Y^m. WIthout loss of generalityn, let us consider the first indirect attribute above. It computes a between x and all neighbours of y in G_X^n. An aggregate function φ from the set $\{min, max\}$ selects, depending on the attribute, the most appropriate value from this set of values. For example, if the attribute a is the number of shared neighbors (resp. shortest distance), φ will be the max (resp. min) function. Notice that both projection parameters n and m can take different values.

D. Link prediction in a projected graph

A projected graph is often seen as a simple unimodal graph (e.g., a co-authorship graph). For predicting the appearance of a new link in a projected graph, learning examples required could then be characterized by a set of attributes computed by direct topological attributes computed in the projected graph only, as already done in [8]. Following our approach, we enrich this set of attributes by a set of other attributes, refered to as *indirect attributes*. Let $a^d_{G_X^n}(x,y)$ be a direct topological attribute computed in the projected graph G_X^n. The d index is used to denote that the topological attribute a is computed

directly in the graph G^n_X which is a unimodal graph (i.e., a projected graph).

Let us suppose that the authors set is represented by the X set in the author-publication bipartite graph. Given two authors (x, y), the common neighbors (denoted CN) of x and y in the co-authorship graph is given by :

$$CN^d_{G^n_X}(x, y) = |\Gamma_{G^n_X}(x) \cap \Gamma_{G^n_X}(y)|$$

The indirect common neighbors attribute is then defined for the same authors $x, y \in G_X$ by :

$$CN^i_{G^n_X}(x, y) = MAX_{u\in\Gamma_G(x),v\in\Gamma_G(y)} CN^d_{G^m_Y}(u, v)$$

The i index is used to denote an indirect attribute. The indirect common neighbor attribute of two authors (x, y) is given by the maximum number of common neighbors of articles signed by each author, computed directly in the dual projected graph G^m_Y, the graph obtained by projecting the bipartite graph G on the publication set Y. More generally, for each direct attribute computed in the X projected graph $a^d_{G^n_X}(x, y)$, we associate an indirect metric given by:

$$a^i_{G^n_X}(x, y) = \varphi_{u\in\Gamma_G(x),v\in\Gamma_G(y)}(a^d_{G^m_Y}(u, v))$$

with the same definition of φ and $\Gamma_G(x)$ as before. Next, we show the effect of introducing these indirect attributes on the performances of the obtained prediction models.

IV. Experimental Results

In order to perfrom a preliminary evaluation of the proposed approach, we have applied it on two different recommendation tasks : an academic recommendation task, where the goal is to predict the occurrence of new links in DBLP bibliographical files. The second application is about recommending music pieces to customers, and has been provided by an industrial partner[3]. For each dataset, we compute predictions by applying a standard *Ada boost* algorithm on a pruned decision tree (J48) with their default values from the Weka system [26]. We apply this supervised learning system using 1) direct topological attributes only and 2) direct and indirect attributes. Performances of learned prediction models are measured in terms of prediction precision, recall and F-measure, and obtained by 10 fold cross-validation. Examples are described in terms of (direct and indirect) following attributes : common neighbours, Jaccard coeffcient, Adamic Adar, preferential attachment (product of nodes' degree), product of nodes' PageRank [18].

A. Academic collaboration recommendation

We have extracted several publications networks form the DBLP bibliographical server. We use publications over a period of 4 consecutive years in order to learn a model. The class labeling is obtained from a network constructed over two years following the end of the learning graph. Three different sets has been used for the moment. Characteristics of these sets, in terms of nodes and links number are given in table I.

[3]Mondomix in the context of the CADI contract http://cadi.lipn.fr

	[1970,1973]	[1972,1975]	[1974,1977]
Authors	2661	4536	6844
Publications	1487	2542	4128
Collaborations	6634	10855	16281
C_c	884	1430	1902
N_c	156	535	1148
D_s	0.00073	0.00044	0.00032

TABLE I
CHARACTERISTICS OF USED DBLP DATASETS AND CORRESPONDING COLLABORATION GRAPH ANALYSIS. N IS THE NUMBER OF NODES IN THE COLLABORATION (PROJECTED) GRAPH, C_c IS THE NUMBER OF CONNECTED COMPONENTS, N_c IS THE NUMBER OF NODES IN THE GIANT COMPONENT AND D_s IS THE GRAPH DENSITY.

	2003	2004
Customers	1212	4614
Items	1217	4630
Transactions	7759	34395

TABLE II
CHARACTERISTICS OF THE BIPARTITE (CUSTOMER,PRODUCT) GRAPH.

For all three datasets, we observe (see figure 3) a clear enhancement, especially in terms of precision of obtained prediction when using indirect topological attributes. Actually, because the co-authorship graph is a projection of a bipartite graph, it lacks a lot of information about the publications. Indirect attributes we introduce allow to compute affinity, or similarity between authors based also on the characteristics of the publication graph, which explains the significant improvement in the performance of the prediction models obtained.

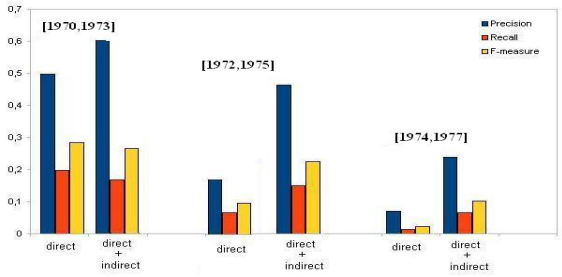

Fig. 3. Performances of obtained prediction models for the DBLP datasets (positive class only).

B. E-commerce recommendation

Similarly, we have extracted two snapshots from the transaction dataset of Mondomix, one for year 2003 and one for year 2004, the characteristics of which are detailed below in tables II and III For each of these snapshots, we have used sales over a period of 6 months in order to build the learning graph and then used the 3 following months to label the pairs (customer,product) as positive or negative. Let us notice that a customer, once identified by the system, may buy as many songs as he wishes. Buying an album is considered as buying all songs contained in the album.

	2003	*2004*
N	1212	4276
C_c	7	33
N_c	940	2908
D_s	0.0439	0.0158

	2003	*2004*
N	1217	4283
C_c	6	23
N_c	870	2790
D_s	0.0426	0.0163

TABLE III
CHARACTERISTICS OF THE *Customer* GRAPH (LEFT TABLE) AND OF THE PRODUCT GRAPH (RIGHT TABLE). N IS THE NUMBER OF NODES, C_c IS THE NUMBER OF CONNECTED COMPONENTS, N_c IS THE NUMBER OF NODES IN THE GIANT COMPONENT AND D_s IS THE GRAPH DENSITY.

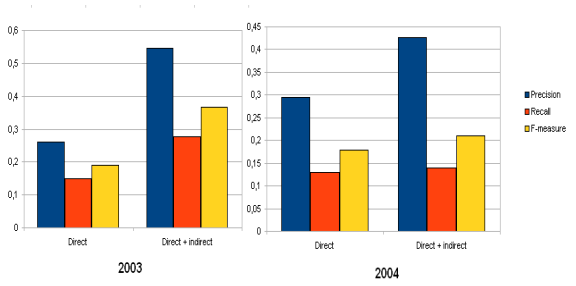

Fig. 4. Performances of obtained prediction models for the Mondomix datasets (positive class only).

For the two studied datasets, we also observe (see figure 4) a clear improvement in the prediction model, also especially in terms of precision of obtained prediction when using indirect topological attributes, taking into account the structure of the projected customer and product graphs to describe each (customer,product) pair.

V. CONCLUSION

In this paper, we have studied the problem of link prediction in a special type of networks obtained by the projection of a bipartite graph over a set of nodes. We have introduced new topological metrics that can reflect the likelihood (or the score) of a link between two nodes that are computed in the dual graph: the graph obtained by the projection of the original bipartite graph but over the other set of nodes. These metrics are used in a dyadic topological supervised machine learning approach for link prediction. We showed, through experimentations over several real world data that new metrics do enhance obtained results, especially in terms of prediction precision, whether the link to be predicted occurs in the original bipartite graph or in one of the projected graph. We have considered here unweighted networks and we have ignored temporal issues when composing the learning graphs. We are working now on adapting the proposed approach in order to handle weighted temporal networks.

VI. ACKNOWLEDGMENTS

This work is part of the LIPS funded by the University Paris Nord as part of its internal projects in 2009. One of the author has been partially funded by the CADI project, with a grant from the French National Agency of Research (ANR) 2008-2010.

REFERENCES

[1] D. Liben-Nowell and J. M. Kleinberg, "The link-prediction problem for social networks," *JASIST*, vol. 58, no. 7, pp. 1019–1031, 2007.
[2] S. P. Borgatti and M. G. Everett, "Network analysis of 2-mode data," *Social Networks*, vol. 19, pp. 243–269, 1997.
[3] S. P. Borgatti, "Social network analysis, two-mode concepts in," in *Encyclopedia of Complexity and Systems Science*, R. A. Meyers, Ed. Springer, 2009, pp. 8279–8291.
[4] Z. Huang, X. Li, and H. Chen, "Link prediction approach to collaborative filtering," in *JCDL*, M. Marlino, T. Sumner, and F. M. S. III, Eds. ACM, 2005, pp. 141–142.
[5] M. E. J. Newman, "Scientific collaboration networks. II. shortest paths, weighted networks, and centrality," *Phys. Rev. E*, vol. 64, no. 1, p. 016132, 2001.
[6] ——, "Co-authorship networks and patterns of scientific collaboration," in *Proceedings of the National academy of science of the uninted state: (PNAS)*, vol. 101, 2004, pp. 5200–5205.
[7] D. Liben-Nowell, "An algorithmic approach to social networks," Ph.D. dissertation, M.I.T., june 2005.
[8] M. A. Hasan, V. Chaoji, S. Salem, and M. Zaki, "Link prediction using supervised learning," in *SIAM Workshop on Link Analysis, Counterterrorism and Security with SIAM Data Mining Conference*, Bethesda, MD, 2006.
[9] M. Pavlov and R. Ichise, "Finding experts by link prediction in co-authorship networks," in *FEWS*, ser. CEUR Workshop Proceedings, A. V. Zhdanova, L. J. B. Nixon, M. Mochol, and J. G. Breslin, Eds., vol. 290. CEUR-WS.org, 2007, pp. 42–55.
[10] T. Murata and S. Moriyasu, "Link prediction based on structural properties of on-line social networks," *New Generation Computing*, vol. 26, pp. 245–257, 2008.
[11] B. Taskar, M. F. Wong, P. Abbeel, and D. Koller, "Link prediction in relational data," in *NIPS*, S. Thrun, L. K. Saul, and B. Schölkopf, Eds. MIT Press, 2003.
[12] A. Popescul and L. H. Ungar, "Structural Logistic Regression for Link Analysis," in *2nd Workshop on Multi-Relational Data Mining (MRDM 2003)*, S. Dzeroski, L. D. Raedt, and S. Wrobel, Eds., August 2003. [Online]. Available: http://www-ai.ijs.si/SasoDzeroski/MRDM2003/
[13] Y. Yamanishi, "Supervised bipartite graph inference," in *NIPS*, D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, Eds. MIT Press, 2008, pp. 1841–1848.
[14] M. Lahiri and T. Y. Berger-Wolf, "Structure prediction in temporal networks using frequent subgraphs," in *CIDM*. IEEE, 2007, pp. 35–42.
[15] M. Berlingerio, F. Bonchi, B. Bringmann, and A. Gionis, "Mining graph evolution rules," in *Machine Learning and Knowledge Discovery in Databases (ECML/PKDD)*. Springer, 2009, pp. 115–130.
[16] R. J. Cooke, "Link prediction and link detection in sequences of large social networks using temporal and local metrics," Master Thesis, University of cape Town, November 2006.
[17] T. Tylenda, R. Angelova, and S. Bedathur, "Towards time-aware link prediction in evolving social networks," *The 3rd SNA-KDD Workshop 09 (SNA-KDD09), June 28, 2009,Paris, France.*, 2009.
[18] S. Brin and L. Page, "The anatomy of a large scale hypertextual web search," in *proceedings of seventh International conference on the world wide web*, 1998.
[19] J. Kleinberg, "Authoritative sources in hyperlinked environments," *Journal of the ACM*, vol. 4, pp. 604–632, 1999.
[20] A.-L. Barabasi, H. Jeong, Z. Néda, E. Ravasz, A. Schubert, and T. Vicsek, "Evolution of the social network of scientific collaboration," *Physica A*, vol. 311, no. 3-4, pp. 590–614, 2002.
[21] S. Wasserman and K. Faust, *Social Network Analysis: Methods and Applications (Structural Analysis in the Social Sciences)*. Camridge University Press, 1994.
[22] L. A. Adamic, O. Buyukkokten, and E. Adar, "A social network caught in the web," *First Monday*, vol. 8, no. 6, 2003.
[23] L. Katz., "A new status index derived from sociometric analysis." *Psychometrika, 18(1)*, vol. 18, no. 1, pp. 39–43, 1953.
[24] F. Fouss, A. Pirotte, J. Renders, and M. Saerens, "Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation," *IEEE. Trans Know. Data Eng.*, vol. 19, no. 3, pp. 355–369, 2007.
[25] C. Faloutsos, K. McCurley, and A. Tomkins, "A. fast discovery of connection subgraphs." in *10th ACM Conference on Knowledge Discovery and Data Mining*, 2004.
[26] I. H. Witten and E. Frank, *Data Mining: Practical machine learning tools and techniques, 2nd Edition*. Morgan Kaufmann, 2005.