

Discovering Collaboration Opportunities in Research-Oriented Networks

Alessandro Cucchiarelli and Fulvio D'Antonio¹

Abstract. In this paper we present a method for discovering collaboration opportunities in *research-oriented networks* based on *opportunity networks*, modelled as a directed attributed multigraph, an opportunity exploiter, a set of opportunity patterns and a set of opportunity ranking functions. We show how the discovery of joint-research collaboration opportunities in a research-oriented network can be formulated in terms of the proposed model, and we perform an evaluation on a real-world dataset.

1 Introduction

Social Networks [5] are constituted by set of actors that are linked together by some kind of social relationship (e.g. friendship, marriage, business collaboration). One of the most interesting aspects of *Social Networks* is the fact that they can also be the source of *opportunities*, defined as situations that can be exploited obtaining valuable outcomes. Opportunity types strictly depend on the nature of social relationships existing in a specific network: examples include finding new friends, meeting a soul-mate or establishing new business collaborations.

In this paper we use the social networks paradigm to model research communities as *research-oriented networks*, whose nodes represent research organizations/industries producing research papers/documents individually or by joint collaborations. The problem we want then to explore is the following: given a research-oriented network, *how can we discover proficient collaboration opportunities?* We provide a solution to this problem using an original model we called *opportunity networks* [2]. Research collaboration is traditionally established on the basis of compatible or complementary research themes but it also heavily depends on human and social aspects, such as mutual trust, previous successful collaborations or political choices. We shall focus our attention only on the former aspect: the assumption is that potentially successful collaboration is based on the similarity of research interests. This ensures a common background knowledge that facilitates mutual understanding therefore speeding up the joint-research production.

2 Opportunity Networks

An opportunity network is a tuple (G, E, O, R) where G is a network modelling the relationships among actors in the domain of interest, represented as a directed attributed multigraph, E is an Exploiter, i.e. an actor interested in discovering opportunities, $O = \{Opp_1, \dots, Opp_k\}$ is a set of Opportunity patterns expressed as graph transformations, and $R = \{Ranking_1, \dots, Ranking_h\}$ is a set of

Ranking functions, where every $Ranking_i : Matches(Opp_i) \rightarrow \mathbb{R}_+$ is a mapping from the set of possible matches for the pattern Opp_i in G to non-negative real numbers. Informally speaking, graph transformations can be modelled as a pair of graphs (L, R) , called left hand side L and right hand side R . Applying the transformation $p = (L, R)$ means finding a match for the graph L in the graph G and substituting L by R obtaining the transformed graph G' . Technically the matching process is called *embedding* and it is carried out by constructing a graph morphism from L to G . It is not the aim of this paper to detail a description of graph transformations concepts (see [3]). Transformations induced in the network can include (in order of increasing complexity) modifying node/attributes, adding new edges, creating cliques, deleting subgroups etc. The *exploiter* is the actor searching for opportunities, in order to obtain valuable outcomes from them. No special constraint is given about the nature of the exploiter: it may be an internal exploiter, i.e. a node of the graph or a sub-graph (a single enterprise or a consortium), or an external exploiter interested but not directly involved in the phenomena occurring in the network (e.g. the case of public administrations “observing” and stimulating the growth of industrial networks). For a detailed definition of Exploiters, Opportunities and Ranking Functions see [2].

3 Research oriented networks

In order to discover joint-research opportunities in a research community we create a graph G with a single node type representing research units and two types of attributed edges: *similarity links* and *co-authorship links*. The creation of the similarity edges involves a number of complex steps:

- given a set RU of research units and a corpus D of documents produced by these units, we annotate such documents with terminological (weighted) vectors of automatically extracted terms [4]. Now we have for each unit r in RU a set of vectors representing a summary of the documents produced by r . We call this set of documents $r(D)$.
- We compute for each $r \in RU$ the centroid $r_{centroid}$ (i.e. the mean vector) of $r(D)$.
- We create a network $G = (RU, E)$ where the nodes are constituted by the research units and the edges belong to $E = E_{sim} \cup E_{coauth}$, where:
 - E_{sim} are edges connecting each pair (a, b) in RU with weight equals to the *cosine similarity*[1] of the two vectors $a_{centroid}$ and $b_{centroid}$. Such edges and the associated weights represent an estimation of the similarity of the research units based on the centroids of the documents produced.

¹ Polytechnic University of Marche, Italy, email:{cucchiarelli, dantonio}@diiga.univpm.it

- E_{coauth} edges are weighted with the number of joint-produced documents for each pair (a, b) of units in RU .

We created a graph G for a real world research network starting with the partner publications on the INTEROP-Vlab web site². The size of this corpus is approximately 1500 papers.

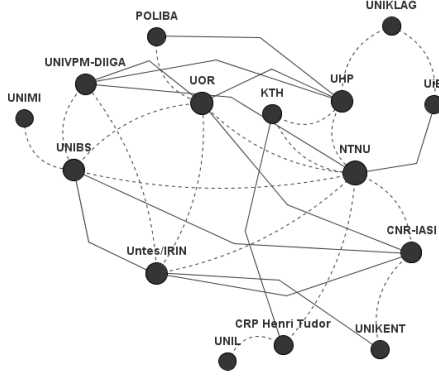


Figure 1. The INTEROP similarity/coauthorship network

In figure 1 a detail of the INTEROP network (obtained deleting edges whose weight is below the mean of all edges' weights) is shown. Curved dashed lines are used for similarity edges and bent lines for coauthorship edges.

The opportunity of a collaboration (*opportunity patterns* R in the transformation $p = (L, R)$) is related to the absence of an edge of co-authorship in the network, i.e. opportunities are expressed by a set of transformations aiming at creating new edges in E_{coauth} . The *exploiter* E (each research unit) takes advantage of the opportunity by creating a new edge. For details on how this pattern can be expressed using graph transformation formalism see [2]. Finally, about the *rankings* R , we assume that new research collaborations are *likely* to be established between two research units a and b if the *similarity* of a and b research interests is *high*, because it means that a and b share a lot of interests. We formalise here this argument by using a single ranking function: $foreach\{x, y\} \in Matches(Opp_1, G)$, $ranking_1(x, y) = w_{sim}(x, y)$. This function expresses the fact that a collaboration opportunity between two research units has a rank that is directly proportional to the weight of the similarity edge associated to them (if any).

4 Evaluation

In this section we evaluate the predictive power of the opportunities proposed by the model. Intuitively our model is able to predict collaborations if the opportunities discovered at a time point t have been established at a subsequent time t' . We do not expect our model to be fully predictive. Our intent is to have a tool that can discover opportunities which might not be picked up by research units, thus acting as a real recommendation system. However, if the model exhibits a good behaviour in predicting collaborations, this is a partial verification of the fact that the opportunities proposed are, in some sense, “valid”.

² <http://interop-vlab.eu/km>

We have partitioned the set of INTEROP documents in 4 subsets based on the year of publication: *papers published before 2004*, *papers published in 2004*, *papers published in 2005* and *papers published after 2005*. We then computed the similarity/coauthorship networks for each subset: $G_{bef2004}$, G_{2004} , G_{2005} and $G_{after2005}$. The coauthorship links that are in a network, but not in the networks that are temporally precedent, are the new collaborations that have been established. In tables 1 and 2 we show the percentage of collaborations (the opportunities) predicted, respectively, for the $G_{bef2004}$ and G_{2004} networks, that have been established by the end of the project.

Year	Exploited Opportunities
2004	20%
2005	33%
after 2005	57%

Table 1. Predictive power based on opportunities in $G_{bef2004}$

Year	Exploited Opportunities
2005	54%
after 2005	75%

Table 2. Predictive power based on opportunities in G_{2004}

The model seems to have a good predictive ability. In particular, the collaborations predicted in the network G_{2004} have higher rate of exploitation with respect to the ones in the network $G_{bef2004}$. This can somehow be justified by the fact that $G_{bef2004}$ comes from papers published before the start of the project and so could fail to represent up-to-date research interests of partners, while the G_{2004} contains the 2004 publications which contribute to define better the research themes currently addressed by them.

5 Conclusions

In this paper we have shown how *opportunity networks* can be applied to implement the task of discovering and ranking joint-research opportunities in a research-oriented network. In future works we will explore more complex opportunity patterns and different application domains, such as discovering business opportunities in industrial or business networks.

REFERENCES

- [1] Ricardo Baeza-Yates and Berthier Ribeiro-Neto, *Modern Information Retrieval*, Addison Wesley, 1st edn., May 1999.
- [2] Alessandro Cucchiarelli and Fulvio DAntonio, ‘Opportunity discovery through network analysis’, in *Proceedings of the International Conference on Interoperability for Enterprise Software and Applications (I-ESA 2010)*, Coventry, UK, (2010).
- [3] *Handbook of graph grammars and computing by graph transformation: volume I. foundations*, ed., Grzegorz Rozenberg, World Scientific Publishing Co., Inc., River Edge, NJ, USA, 1997.
- [4] Francesco Sclano and Paola Velardi, ‘Termextractor: a web application to learn the shared terminology of emergent web communities’, in *Proceedings of the 3rd International Conference on Interoperability for Enterprise Software and Applications (I-ESA 2007)*, Funchal, Portugal, (2007).
- [5] Stanley Wasserman, Katherine Faust, and Dawn Iacobucci, *Social Network Analysis : Methods and Applications (Structural Analysis in the Social Sciences)*, Cambridge University Press, November 1994.