Experimental Evaluation of Academic Collaboration Recommendation Using Factorial Design

Michele A. Brandão, Mirella M. Moro and Jussara M. Almeida

Universidade Federal de Minas Gerais, Brazil
{micheleabrandao, mirella, jussara}@dcc.ufmg.br

Abstract. Recommender systems have been used in e-commerce and online social networks. Among various challenges to construct such systems, how to parameterize them and their evaluations are two vaguely explored issues. Generally, each recommendation strategy has parameters and factors that can be varied. In this article, we propose to evaluate the impact of key parameters of two state-of-the-art functions that recommend academic collaborations. Our experimental results show that the factors affect recall, novelty, diversity and coverage of the recommendations in different ways. Finally, such evaluation shows the importance of studying the impact of the factors and factor interactions in the academic collaboration recommendations context.

Categories and Subject Descriptors: H.3 [**Information Systems Applications**]: Collaborative and social computing systems and tools; H.5 [**Information Retrieval**]: Evaluation of Retrieval Results

Keywords: collaboration recommendation, factorial design, factors

1. INTRODUCTION

Recommender systems have been used in e-commerce and online social networks, among others [Brandão et al. 2013b; Freyne et al. 2010; Lopes et al. 2010]. Besides the great success of online social networks that promote friend relationships (e.g. Facebook) or work relations (e.g. LinkedIn), there are also academic social networks in which the nodes represent the researchers and the edges their co-authorships. In such context, the collaboration recommendation among researchers is relevant because it may help a researcher to form new groups or teams, to search for collaborations when writing a grant proposal, to improve the quality of communication in the network and to investigate different research communities [Brandão et al. 2013b]. Also, a recent work shows that research groups with a well connected co-authorship social network tend to be more prolific [Lopes et al. 2011]. Moreover, collaboration is normally a good thing from a wider public perspective [Adams 2012], and good connections within the network are critical to learning and creativity [Burt 2004].

After establishing a recommendation function, another important issue is how to evaluate it. Recommender systems may be evaluated by user feedback mechanisms. However, in many contexts (such as academic or peer recommendation), giving feedback is complex because a person might assess a recommendation as "bad" due to subjective matters, such as personally not liking the one recommended, lack of affinity or even competition. Another option is to evaluate recommendations based on user preferences (affinity with other researchers or lack of it) by modeling their behavior (collaborations regarding their preferences) and using such model to evaluate the recommendations [Shani and Gunawardana 2011]. This strategy is also not appropriate in the academic setting, because there is no public data with preference information of a researcher in relation to others.

Therefore, collaboration recommendations have been evaluated by splitting the academic social

This work was partially funded by CAPES, CNPq, Fapemig and InWeb, Brazil.

network in two parts [Brandão et al. 2013b; Lopes et al. 2010]. The first split is used to apply the recommendation functions and to generate a recommendation list. The second split is used to establish a comparison between the collaborations effectively made during that period (*ground truth*) and the recommendations. Then, the recommendation evaluation verifies whether the recommended collaborations are present in the *ground truth*. Note that this strategy does not influence the recommendations results for a final user, but influences the results in the evaluation itself.

Among various challenges to generate collaboration recommendations, how to parametrize the recommendation function as well as the adopted evaluation strategy (regarding data splitting) are still open issues. In other words, it is important to analyze how the variation in the size of the splits as well as in the values of key parameters of the functions impact the results of the collaboration recommendation. Specifically here, we consider two state-of-the-art collaboration recommendation functions: *Affin* by Brandão et al. [2013b] and *CORALS* by Lopes et al. [2010]. These functions are the only collaboration recommendation functions in the literature that combine different weights to represent link semantic among pairs of researchers. Such weights may impact the resulting recommendations; however, there is no theoretical nor experimental analysis of the sensitivity of the recommendation functions regarding such parameters and their possible interactions.

This work aims to analyze the impact of the defined data splitting as well as the weights explored by the aforementioned recommendation functions. This analysis consists of performing a 2^k factorial experimental design in which each parameter upon analysis (called *factor*) can take two values (or levels). This design allows the evaluation of the relative impact of each factor as well as of the interaction of multiple factors on the response variable being studied (in our context, the efficacy of the recommendations) [Jain 1991]. Note that there is interaction between factors when the effects of a factor in the response variable depends on the value of the other. Here, we consider a 2^k design with $k = 3$ factors: the splitting strategy of the total number of publications, and the two weights explored by the recommendation functions. We perform separate experiments for the *Affin* and *CORALS* functions in order to study the relative impact of each factor on the effectiveness of each function. After performing the factorial design and identifying the factors that have more impact on each response variable, we further evaluate the most important factor, notably the splitting strategy by varying it at a finer granularity (as opposed to only two levels as in the 2^k design).

The collaboration recommendations are evaluated regarding four metrics: recall, novelty, diversity and coverage, which are suitable for such a context as described by Brandão et al. [2013b]. Among a large number of metrics discussed in [Shani and Gunawardana 2011], these are more appropriate to evaluate recommendation of collaborations. Other metrics such as confidence, trust and utility are not appropriate, because prior information about researchers' preferences is necessary, but beyond our reach. We also consider the total number of recommendations and the number of correctly returned recommendations as evaluation metrics, because they relate to recall. In other words, for each recommendation function and each of these six metrics, we perform a 2^k factorial design. The analysis performed with the factorial design shows that the factors impact the metric results in different ways. Then, we perform experiments for each response variable impacted by the splitting strategy (alone or in interaction). These evaluations show the potential of using the factorial design in the evaluation of recommender systems and other applications dependent on the parameter setting data.

After presenting a study of the related work (Section 2), the contributions of this article are: a discussion of the academic collaboration recommendation functions of the state-of-the-art (Section 3); a description of the factors and specification of factorial design (Section 4); an identification and analysis of the impacting factors (Section 5.1) and a detailed analysis of the impact of the splitting strategy on the response variables (Section 5.2). Finally, we conclude this article and discuss possible directions for future work in Section 6.

2. RELATED WORK

In this section, we discuss related work on recommendations in social networks and experimental design. Then, we emphasize the contributions of our work in the presence of such state-of-the-art.

Social networks and recommendations. Social networks is a very prolific research area within Computer Science. Examples of problems include the prediction of links in networks using different approaches [Cohen and Cohen-Tzemach 2013; Kuo et al. 2013] and the friend recommendation on an online social network [Guimarães et al. 2013]. There are also several efforts to develop methods to recommend items and people on social networks. For instance, Freyne et al. [2010] and He and Chu [2010] recommend items by exploring user preferences and interactions. On the other hand, people recommendation needs to consider aspects of social connections [Guimarães et al. 2013; Lopes et al. 2010]. In the specific context of academic social networks, Brandão et al. [2013b] and Lopes et al. [2010] present new collaboration recommendation functions (which are briefly described in Section 3). Both have evaluated the proposed functions under fixed parameter (weights) settings and have not studied how the values of such parameters impact the effectiveness of the methods, which is the aim of this work. The recommendations can be evaluated with an online strategy, in which users assess how good the recommendations are, or offline, in which the recommendations are compared with a ground truth without user intervention [Shani and Gunawardana 2011]. For example, the online strategy has been used by Netflix and Amazon, whereas the offline approach has been applied to evaluate the recommendation academic collaborations [Brandão et al. 2013b; Lopes et al. 2010].

Experimental design. Different experimental designs can be applied to solve distinct problems. A proper analysis of experiments depends on the choice of such a design. Here, we perform a 2^k factorial design without replication [Jain 1991] to quantify the impact of different factors on the recommendations produced by the functions proposed by Brandão et al. [2013b] and Lopes et al. [2010]. The factors are the offline evaluation strategy (splitting the social network) and specific parameters of the two recommendation functions. The 2^k factorial design has been applied in various contexts, such as a guideline to parametrize genetic programming algorithms [Lima et al. 2010] and to investigate the influence of various factors on the synthesis of single-core superparamagnetic iron oxide nanoparticles [Lak et al. 2013]. In a 2^k factorial design, each factor (parameter) under study is evaluated under two levels (parameter values), and all 2^k combinations of factor levels are considered.

A more general design is a full factorial, where each parameter is evaluated under an arbitrary number of levels. It uses every possible combination of all factor levels, i.e. every possible combination of configuration is analyzed. However, the number of experiments required for a full factorial design is often too large (specifically, when the number of factors or their levels is large). In such cases, one can use only a fraction of the full factorial design. Fractional factorial designs save time and money when compared to full factorial designs, but they also provide less information than a full factorial design. Both types of designs have been applied in different contexts. For example, [Brumec and Vrček 2013] used fractional factorial design for comparing the costs of leasing IT resources and Buragohain and Mahanta [2008] applied a full factorial design for I/O optimization in training neural networks. Together with the simple design (where each factor is studied in isolation), these three are the most used designs [Jain 1991]. Here, we adopt a 2^k factorial design as well as a simple design to study the impact of various parameters on the collaboration recommendation functions.

Discussion on contributions. Different recommendation approaches have been developed by focusing on distinct types of social networks including recommendation in academic social networks, which is our context. Likewise, different techniques of experimental design have been applied in various scenarios. In this article, we apply some of these techniques (2^k factorial design and simple design) to study the impact of parameters of state-of-the-art recommendation functions on academic collaborations. The main difference of our work is applying experimental designs to study parameters that may impact on collaboration recommendations. In [Brandão et al. 2013a], we applied a 2^k factorial design to quantify the impact of the parameters of *Affin* and *CORALS* collaboration recommendation functions. Here, we build on it by extending our evaluation to further study the impact of one of the most relevant factors identified by the 2^k design: we focus on the splitting strategy, and apply a simple experimental design to evaluate how it impacts the effectiveness of the recommendations at a much finer granularity.

3. RECOMMENDATION FUNCTIONS OF ACADEMIC COLLABORATIONS

Collaboration recommendation is a specific recommendation problem in which two individuals are recommended to work together. In order to achieve relevant recommendations, it is necessary to consider aspects that influence collaboration relationships. For instance, in *CORALS* (*Collaboration Recommendation on Academic Social Networks*) [Lopes et al. 2010] and *Affin* [Brandão et al. 2013b], a weight represents each relation among researchers, which then is combined with other information within recommendation functions. The recommendation function returns a list of recommended pairs of researchers to initiate collaboration. The *Affin* and *CORALS* recommendation functions recommend pairs of researchers i and j to collaborate according to Equations 1 and 2, respectively.

$$r_{i,j} = \begin{cases} Initiate, & \text{if } (Cp_{i,j} = 0) \wedge \\ & (Affin\_Sc_{i,j} > threshold); \end{cases} \quad (1)$$

$$r_{i,j} = \begin{cases} Initiate, & \text{if } (Cp_{i,j} = 0) \wedge \\ & (Cr\_Sc_{i,j} > threshold); \end{cases} \quad (2)$$

$Cp_{i,j}$ measures how a researcher i has collaborated with another researcher j, where *zero* indicates that such pair has yet not collaborated. $Affin\_Sc_{i,j}$ (Equation 1) is a weighted average between *Affin* and social closeness Sc, in which *Affin* quantifies how much a researcher i has co-authored with people from j's institution, and Sc measures the shortest path between pairs of researchers i and j in the co-authorship social network. $Cr\_Sc_{i,j}$ (Equation 2) is a weighted average between correlation Cr and social closeness Sc, in which Cr quantifies how much the pair of researchers i and j have published in similar areas of research.

The function *Affin* recommends pairs of researchers to initiate collaboration when $Cp_{i,j}$ is zero and $Affin\_Sc_{i,j}$ is greater than a predefined threshold. This threshold represents the minimum value of the weighted average among researchers relations and is defined according to ranges that may follow a linear scale, such as low $< 33\%$ and high $> 66\%$. We choose "low" as threshold. On the other hand, *CORALS* recommends to initiate collaboration when $Cp_{i,j}$ is zero and $Cr\_Sc_{i,j}$ is greater than "low" (chosen threshold). The threshold is also a parameter of the recommendation functions, but it was not included in the 2^k factorial design because the individual weights (social closeness, affiliation and correlation) have a greater impact on the academic collaboration recommendation.

4. EXPERIMENTAL DESIGN

The main purpose of an experimental design is to obtain the maximum amount of information with the minimum number of experiments [Jain 1991]. As aforementioned, there are various types of experimental designs, such as simple design, full factorial design, and so on. In this work, we perform a 2^k factorial design, in which k represents the number of factors and 2 the number of levels that each factor has[1]. The factors are the parameters that affect the results of the experiments, and the levels are the values that each factor may take. A 2^k factorial design was chosen for allowing the study of the impact of the factors and factor interactions in the response variables. It is important to note that we have performed the factorial design without replication. In other words, only one result is produced for each configuration defined by the factor levels. This was done because the recommendation function algorithms are deterministic[2]. Besides the 2^k factorial design, we have also performed a simple design to evaluate the impact of the splitting strategy factor in some response

[1] Here, the two levels of each factor are called upper and lower levels.
[2] In the case of a random process, the 2^k factorial design with replication could be adopted.

Table I: Splitting Strategy

		First Split		Second Split	
Description	**Level**	**% Data**	**# Publication**	**% Data**	**# Publication**
lower	1	10%	1.386	90%	12.481
	2	**20%**	**2.773**	**80%**	**11.094**
	3	30%	4.160	70%	9.707
	4	40%	5.546	60%	8.321
equal	5	50%	6.933	50%	6.933
upper	6	60%	8.321	40%	5.546
	7	70%	9.707	30%	4.160
	8	**80%**	**11.094**	**20%**	**2.773**
	9	90%	12.481	10%	1.386

Table II: Levels Definition

Factor	**Level**	**Value**
Splitting Strategy	lower	20% - 80%
	upper	80% - 20%
Weights	lower	10
	upper	100

variables at a finer granularity, that is, considering more than 2 levels. Therefore, we here present the dataset and detail the application of the 2^k factorial design and simple design in our context.

Our experiments use real data from the DBLP[3] digital library. The academic social network built from DBLP has 629 researchers of 45 Brazilian institutions and their 13,867 publications in journals and conferences dated from 1973 up to 2012. In such network, the nodes are the authors (researchers) and the edges (links between authors) represent co-authorships in publications. Furthermore, since the dataset is divided into two splits, an academic social network is built for each split. The first split has papers published before those on the second split and is used to generate the recommendations. The second split is used to evaluate the resulting recommendations.

For the 2^k design, the factors that can impact the recommendations are: the splitting of the total number of publications in two parts and the recommendation function weights, as defined next.

Splitting strategy. The social network is divided into two different splits, and each possible splitting represents a possible level for this factor, as shown in Table I. For example, in the second level, 20% of the data are used to generate recommendations and contains papers published before the 80% of the data in the second split. The first split is explored to create the researchers' profile and the social network. The second split contains the expected results a recommender system should provide. Furthermore, both splits also follow the time interval distribution, where the first split considers publications prior to the second one. In other words, the second split represents the "future" of the first one, and hence allows to identify which recommendations would be more useful. In this case, a shorter time interval is considered to generate the recommendations. In the eight level, 80% of the data are used to recommend collaborations and 20% of the data to validate them. In order to perform the 2^k factorial design, we have chosen to avoid extreme values, having the 2 and 8 levels as representatives of the upper and lower levels, as shown in Table II. In this case, the extreme level is 90%-10% and choosing it may artificially change the factor effect. Furthermore, we cannot choose intermediate values, because they will not help us decide if the difference in performance (considering the evaluation metrics) is significant enough to justify detailed examination [Jain 1991]. We also study the impact of the splitting strategy at a finer granularity by considering more levels and applying a simple design, as discussed in Section 5.2.

Recommendation weights. The recommendation functions have associated weights that quantify the relationships between pairs of researchers. *Affin* has affiliation and social closeness weights, and *CORALS* has correlation and social closeness weights (social closeness weight is common to both). Each weight represents a factor that may impact the recommendation functions. These factors are numeric and can have different levels. The adopted values as lower and upper levels for each weight are in Table II. These values are chosen because prior experiments show that they provide more stable results (as fully detailed in [Brandão 2013]).

Factorial design. A 2^k factorial design is performed for each function and with three factors (splitting strategy, social closeness weight and correlation/affiliation weight), i.e., $k = 3$ being required $2^3 = 8$ experiments. These factors are represented by variables x_A, x_B and x_C. Each variable takes either -1

[3]DBLP: http://www.informatik.uni-trier.de/~ley/db

or $+1$ values to represent the lower and upper levels of the factor, respectively. Intuitively, the idea behind the design is that the value of the response variable can be regressed on the variables x_A, x_B and x_C using a non-linear additive model shown in Equation 3.

$$y = q_0 + q_A x_A + q_B x_B + q_C x_C + q_{AB} x_A x_B + q_{AC} x_A x_C + q_{BC} x_B x_C + q_{ABC} x_A x_B x_C \tag{3}$$

where y represents the response variable of the factorial design (recall, novelty, diversity, coverage, total of recommendations and correct recommendations), q_0 is the average behavior of the recommendation function independent of factor levels, and each other q_* is the effect of a factor (or factor interaction) in the response variable y. Specifically, $q_0 = \frac{1}{2^3}(y_1 + y_2 + y_3 + y_4 + y_3 + y_4 + y_5 + y_6 + y_7 + y_8)$, where each y_i is the value of the response variable at a specific configuration of parameter levels. For example, y_1 is obtained when the three factors are in the lower levels (i.e., $x_A = x_B = x_C = -1$). The model parameter q_A is the effect of factor A in y, and q_{AB} is the effect of the interaction between A and B factors (Jain [1991] describes the calculation of these effects). The effect of the factors and their interactions can be positive or negative in the response variables (a positive effect indicates positive correlation, whereas a negative effect indicates negative correlation). The effects are used to calculate the percentage of variation of the measured data (i.e., values of y_i) that is explained by each factor. The percentage of variation captures the importance of each factor to the response variable. It is important to note that we consider the factors as important when the percentage of variation explained by them is higher than 10%. Another threshold can be chosen, depending of the application and cost to analyze the factors.

Simple design. In a simple design, the evaluation starts with a typical configuration, and one factor is varied at a time to see how it affects the performance [Jain 1991]. For example, if the affiliation weight factor were the only factor to (mostly) influence the recall (response variable) of the recommendation function *Affin*, we could fix the values of the two other factors - splitting strategy and social closeness - and vary the values of affiliation weight factor at a much finer granularity than the two levels adopted in the 2^k design. Therefore, this design is suitable for factors whose interaction with other factors can be neglected, i.e., the percentage of variation explained by the interaction is too low or null.

Response variables. The collaboration recommendations are evaluated regarding recall, novelty, diversity, coverage, number of correct recommendations and number of total recommendations. Recall measures the fraction of relevant instances that are retrieved. Hence, the higher the recall, the better the results. The novelty metric aims to quantify the "novel" characteristic in a recommendation list [Fouss and Saerens 2008] and is measured based on the frequency that a researcher appear in the recommendations list of other researchers. This frequency represents the popularity degree of the researchers, i.e., researchers with high frequency are likely to be known (low value, high novelty). In this case, we consider that the less popular a recommended researcher, the most probable he/she is unknown to a target researcher. The diversity in a recommendation list is measured by using the intra-list similarity metric [Shani and Gunawardana 2011], where high values indicate low diversity. Coverage is represented by a metric that computes how unequally different the recommended items are to users [Shani and Gunawardana 2011]. Here, we compute such metric through equation based on *Gini index*, as presented in [Shani and Gunawardana 2011]. Correct recommendations measures the number of recommendations returned by recommendation functions that are present in the *ground truth* (higher is better). Finally, the total recommendations quantifies the total number of recommendations (present in the ground truth or not) returned by recommendation functions. In recommender systems, the lower the total recommendations and the higher the correct recommendations, the better the results.

5. EXPERIMENTAL RESULTS

In this section, we present an analysis of the results from applying the 2^k factorial design (Section 5.1) and a detailed analysis of the impact of the splitting strategy (Section 5.2). One problem with 2^k factorial designs is that it is not possible to estimate experimental errors since no experiment is

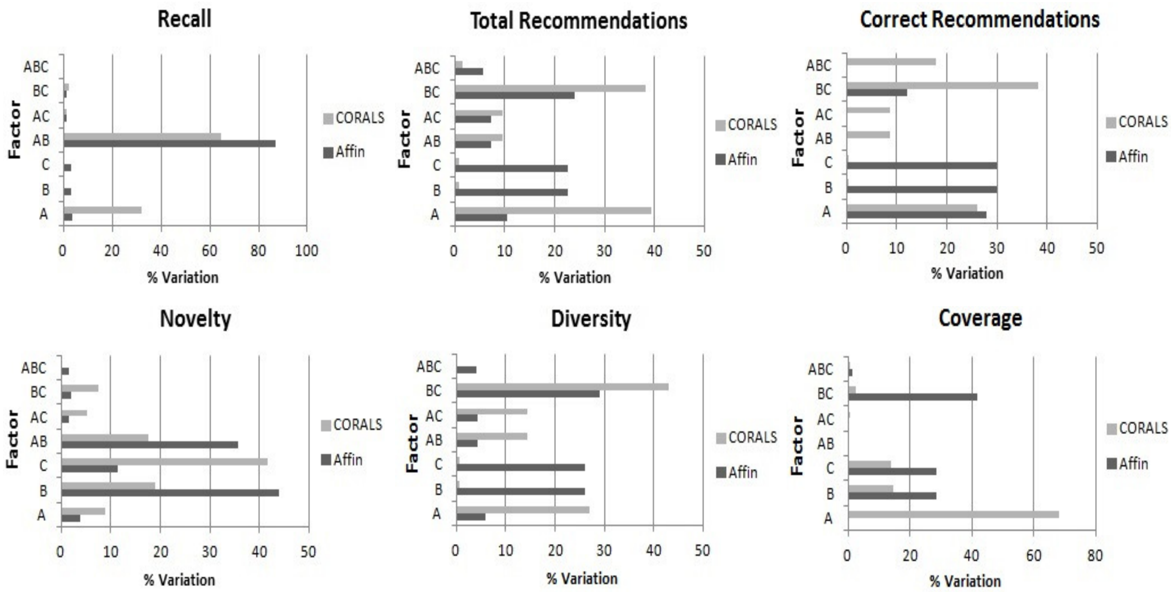

Fig. 1: Percentage of variation explained by each factor on the response variable - *Affin* versus *CORALS*, where A is splitting strategy, B is social closeness weight, and C is affiliation/correlation weight.

repeated. Experimental errors can be quantified by repeating the measurements under the same factor-level combinations [Jain 1991]. Since the recommendation function algorithms are deterministic, it is not possible to perform statistical analysis in this work.

5.1 Results of the 2^k Factorial Design

We quantify the relative impact of the factors on each recommendation function (*Affin* and *CORALS*), and discuss the results for each function separately first. We then make a comparative analysis.

Recommendation function *Affin*. Figure 1 shows the fraction of the total variation observed in each response variable considered - recall, novelty, diversity, coverage, total number recommendations and number of correct recommendations - explained by each factor and factor interaction. The steps to compute these fractions are the following: first, we compute the effect of each factor and factor interaction (q_* mentioned in Section 4); second, we calculate the variation of each factor and factor interaction, and the total variation of all factors and interactions; then, we obtain the fraction of variation dividing the variation of each factor/interaction by the total variation. The calculation of the effects and variations are fully described in Jain [1991].

Also in Figure 1, the interactions between the factors splitting strategy (A) and social closeness weight (B) are responsible for most of the variation in recall (over 80%). The factors social closeness weight (B), affiliation weight (C) and their interaction BC are significant for the total of recommendations; whereas the splitting strategy (A), social closeness weight (B), affiliation weight (C) and their interaction ABC are the most important factors for the number of correct recommendations. For the novelty variable, the factors social closeness weight (B), affiliation weight (C) and AB interaction explain almost all the variation (91.08%). On the other hand, the diversity and coverage are mainly affected by social closeness weight (B), affiliation weight (C) and their interaction BC.

In all response variables, note that at least one two-factor interaction - either the interaction between the splitting strategy and social closeness weight (AB) or the interaction between the two weights (BC) - has significant impact on each considered response variable. This interaction exists when the impact of one factor depends on the level of the other. For instance, the interaction AB explains 86.88% of the variation in the recall of the recommendations. This implies that the splitting strategy (A) should not be considered separately from the social closeness factor (B), because the impact of either one depends on the specific levels of both. In other words, the impact of the splitting strategy on recall depends strongly on the value assigned to the social closeness weight. If the level of social closeness weight changes, the impact of the splitting strategy may change as well.

In contrast, the interaction between the splitting strategy and the affiliation weight (AC) as well as the three-factor interaction have negligible impact on *Affin*, in any of the considered metrics.

Recommendation function *CORALS*. Figure 1 also shows the percentage of the total variation of each considered response variable produced for the *CORALS* function that is explained by each factor and interaction. Note that these percentages are computed the same way as in *Affin*. The splitting strategy (A) as well as its interaction with the social closeness weight (AB) explain almost all the variation on recall. The splitting strategy and the interactions between the two weights (BC) are the most relevant factors for the total of recommendations, whereas the splitting strategy, the interaction between the function weights (BC) and the interaction among all three factors (ABC) explain almost all the variation observed in the number of correct recommendations. Regarding the variations on novelty, the social closeness weight (B), the correlation weight (C) and interaction AB have most impact, whereas the splitting strategy and all two-factor interactions (AB, AC and BC) have significant impact on diversity. Finally, the factors splitting strategy (A), social closeness weight (B) and correlation weight (C) explain most of the variation in coverage. The lack of significant impact of any interaction on coverage implies that the three factors can be evaluated independently from each other in the analysis of this response variable.

Finally, it is also important to note that performing a 2^k factorial design requires the validation of some assumptions. One such assumption is that the effects of different factors (and factor interactions) are *additive* [Jain 1991]. The derivation of the model shown in Equation 3 relies on this assumption. A common mistake is to perform a 2^k factorial design using factors whose effects are not additive (e.g., they are multiplicative). In this case, one possibility is to apply a transformation on the data before doing the factorial design. In case of multiplicative factor effects, the use of a logarithm transformation is suggested such that the additive model (Equation 3) can be applied to the transformed data[4] [Jain 1991]. We emphasize that all assumptions were validated in our experiments. In order to verify whether the assumption of additive factor effects could be influencing our results, we have also performed factorial designs on the data after a logarithm transformation, which produces a multiplicative model, obtaining qualitatively similar results. To that end, we performed 2^2 factorial design with only two factors at a time, fixing the third factor. In each scenario, we performed two factorial designs, one with an additive model and one with a multiplicative model (applying the transformation). The results for both models were very similar implying that the assumption of additive effects is reasonable in our context. Therefore, our results are consistent.

Comparative analysis. Table III shows the most important factors that explain the variations in the response variables and the total percentage of variation caught by those factors. Note that all percentages are greater than 69% and, half of them are over 90%. In other words, the factors presented in Table III explain most of the variations in the response variables.

A comparative analysis of the most important factors for the response variables in each recommendation indicates similarities across functions. Specifically, the interaction between splitting strategy and social closeness weight (AB) significantly influences the recall of both recommendation functions. Furthermore, the total number of recommendations, the diversity and the coverage of *Affin* are not very affected by the splitting strategy adopted (neither isolated nor interacting with any other factor). On the other hand, the effectiveness of *CORALS*, in terms of all response variables considered, are significantly impacted by such factor. Consequently, the evaluation of the collaboration recommendation

[4] This happens because if $y = a * b$, then $log(y) = log(a) + log(b)$.

Table III: Impacting Factors and Percentage of Total Variation (%)

Response Variable	***Affin***	***CORALS***
Recall	AB: 86.88	A/AB: 96.47
Total recommendations	B/C/BC: 69.37	A/BC: 77.59
Correct recommendations	A/B/C/BC: 99.81	A/BC/ABC: 82.23
Novelty	B/C/AB: 91.08	B/C/AB: 78.38
Diversity	B/C/BC: 81.32	A/AB/AC/BC: 98.89
Coverage	B/C/BC: 98.55	A/B/C: 96.24

through social network splitting into two parts influences the effectiveness of the recommendation, particularly of *CORALS*. Furthermore, *CORALS* is more sensitive to the evaluation strategy considered than *Affin*, i.e., it is necessary to pay more attention to the configuration of the splitting strategy when using *CORALS*. This difference between the functions is due to the affiliation and correlation weights, because such weights differentiate the collaboration recommendations made by each function.

5.2 The Impact of the Splitting Strategy

In this section, we further evaluate the impact of the splitting strategy and/or its interaction with the weights of the recommendation functions on the effectiveness of the recommendations. Unlike in the previous section, where we evaluated each factor at only two levels, here we study the impact of the splitting strategy at a finer granularity, considering more levels. We choose to focus on the splitting strategy factor because the 2^k design revealed that it is one of the most important ones to both recommendation functions. Specifically, according to the results presented in Section 5.1, the splitting strategy factor and/or its interactions with the other factors have key impact on recall, number of correct recommendations and novelty in *Affin* and *CORALS* as well as on the diversity and coverage of *CORALS*. In this evaluation, we apply a simple design on varying the splitting strategy at multiple levels. For cases where its interaction with another factor was found to be important, we also perform such fine grain evaluation considering multiple levels of the other factor as well.

We start by focusing on recall. Table IV shows the results for the function *Affin* obtained with different values of splitting strategy. The values selected reflect scenarios where the first split is larger than the second one (level 6 to 9 of Table I), because the effect of such factor (q_A) is positive. In other words, the recall increases with the selected level. Since the interaction of the splitting strategy with the social closeness weight (AB) revealed to be very important to both recommendation functions, we consider two values for the social closeness weight Sc (10 and 100) and, for each such value, vary the splitting strategy in four different scenarios. The affiliation weight is fixed as 10 because we found this factor to be irrelevant, and thus, the value assigned to it does not significantly impact the recall. The results in Table IV shows that when the social closeness weight is set to 10, there is not a clear pattern in how the recall behaves as we increase the first split. The differences in recall across the four considered splitting strategies are somewhat small (up to 23%). The highest recall is achieved with a 70%-30% splitting. However, the recall results greatly improve when the social closeness weight is set to 100, for all splitting strategies. In this case, we see a clear pattern: the recall tends to increase as the first split increases. Moreover, we observe a much greater variation on recall as we vary the splitting strategy (up to 66%), reflecting a greater impact of this factor on recall. The best overall result is for a 90%-10% splitting strategy and a social closeness weight equal to 100.

Table IV also presents recall results for the recommendation function *CORALS* as we vary the splitting strategy and social closeness weight at the same levels. The correlation weight is fixed as 10. In general, the results are similar to those observed for *Affin*: *CORALS* has greater recall when splitting the social network in 90% - 10% and assigning 100 for social closeness weight. Note however that, unlike *Affin*, here we see the same increasing pattern in recall as we increase the first split regardless of the value set to the social closeness weight.

As aforementioned, 2^k design results showed that the impact of the splitting strategy on the number

Table IV: Recall: *Affin*/Cr = 10

Sc \ Split	60%-40%		70%-30%		80%-20%		90%-10%	
Function	*Affin*	*CORALS*	*Affin*	*CORALS*	*Affin*	*CORALS*	*Affin*	*CORALS*
10	0.1407	0.212	0.164	0.263	0.1444	0.33	0.1333	0.4
100	0.3216	0.3367	0.355	0.382	0.4	0.44	0.5333	0.633

Table V: *Affin* - Correct: $Sc = 100$ e *Affin* = 10 and *CORALS* - Total: $Sc = 100$ e $Cr = 10$

Split	Correct Recommendations	Total Recommendations
10% - 90%	39	685
20% - 80%	48	2717
30% - 70%	47	1261
40% - 60%	51	3604
60% - 40%	64	6961
70% - 30%	54	7015
80% - 20%	36	10713
90% - 10%	16	9253

Table VI: *CORALS* - Correct Recommendations

Split=60%-40%

Cr \ Sc	10	100
10	42	67
100	49	42

Split=70%-30%

Cr \ Sc	10	100
10	40	58
100	44	40

Split=80%-20%

Cr \ Sc	10	100
10	30	40
100	29	30

Split=90%-10%

Cr \ Sc	10	100
10	12	19
100	12	12

of correct recommendations produced by *Affin* is independent of the other factors. Thus, we further evaluate the impact of this factor by fixing the values assigned to both weights, and varying the splitting strategy. Specifically, we set the social closeness weight to 100 and the correlation weight to 10, and vary the splitting strategy considering all scenarios specified in Table I. Table V shows that more correct recommendations are returned when the splitting strategy is 60% - 40%. On the other hand, in *CORALS*, the number of correct recommendations is significantly impacted by the interaction among three factors: splitting strategy, social closeness weight and correlation weight. Thus, we further evaluate the impact of the splitting strategy on this response variable considering different combinations of values for the function weights. The results, shown in Table VI, indicate that the number of correct recommendations is greater when the splitting strategy is 60% - 40%, the social closeness weight is 100 and the correlation weight is 10. In other words, a large social closeness weight, a small correlation weight and an intermediary splitting strategy leads to the largest number of correct recommendations by *CORALS*.

The splitting strategy also impacts the total number of recommendations produced by *CORALS*. As this factor does not interact with any of the function weights, we perform a simple design and fix the social closeness and correlation weights in 100 and 10, respectively. The results in Table V show that fewer collaborations are recommended when the splitting strategy is 10% - 90%. It is important to note that the smaller the total number of recommendations returned, the better the result. Despite being a difficult task, the ideal is that the recommendation function increases the number of correct recommendations returned without increasing the total number of recommendations. In order to verify if these response variables are related, we calculate the correlation coefficient (CC) [Rodgers and Nicewander 1988] between them (considering the values produced by the 2^k factorial design). The results show a moderate correlation ($CC = 0.536$) between these response variables for the recommendation function *Affin* and a very weak correlation ($CC = 0.181$) for *CORALS*. The weaker correlation for *CORALS* seems to imply that this function is able to more often raise the number of correct recommendations without necessarily penalizing the total number of recommendations.

Regarding the novelty in the collaborations recommended by *Affin*, we perform further experiments by varying the splitting strategy and the social closeness weight, and fixing the affiliation weight as 10. Table VII shows that higher novelty (lower value, higher novelty) when splitting strategy is 10%

Table VII: *Affin* - Novelty: *Affin* = 10

Split \ Sc	10	100
10% - 90%	0.044	0.026
20% - 80%	0.038	0.05
30% - 70%	0.080	0.032
40% - 60%	0.048	0.058
60% - 40%	0.042	0.043
70% - 30%	0.049	0.044
80% - 20%	0.048	0.06
90% - 10%	0.043	0.070

Table VIII: *CORALS* - Novelty: Cr = 10

Split \ Sc	10	100
10% - 90%	0.031	0.034
20% - 80%	0.029	0.056
30% - 70%	0.092	0.038
40% - 60%	0.031	0.067
60% - 40%	0.024	0.056
70% - 30%	0.026	0.048
80% - 20%	0.032	0.082
90% - 10%	0.035	0.094

Table IX: *CORALS* - Diversity

Split=60%-40%

Cr \ Sc	10	100
10	373.76	582.30
100	747.29	373.76

Split=70%-30%

Cr \ Sc	10	100
10	359.98	549.54
100	795.71	359.98

Split=80%-20%

Cr \ Sc	10	100
10	489.14	1136.89
100	736.27	489.14

Split=90%-10%

Cr \ Sc	10	100
10	16.02	1066.97
100	451.17	388.21

Table X: *CORALS* - Coverage

Split=60%-40%

Cr \ Sc	10	100
10	0.57	0.54
100	0.54	0.57

Split=70%-30%

Cr \ Sc	10	100
10	0.55	0.55
100	0.54	0.55

Split=80%-20%

Cr \ Sc	10	100
10	0.54	0.51
100	0.54	0.54

Split=90%-10%

Cr \ Sc	10	100
10	0.52	0.52
100	0.53	0.52

- 90% and social closeness is 100. For the recommendation function *CORALS*, the same experiments were performed to evaluate the novelty, but fixing the correlation weight in 10. Table VIII shows that the highest novelty is achieved when splitting strategy is 60% - 40% and social closeness weight is 10.

Next, we further evaluate the impact of the three factors on the diversity of *CORALS*. Recall that the three-factor interaction has significant impact on this response variable. We perform experiments varying them all. The results in Table IX show high diversity (lower value) when the splitting strategy level is 90%-10% and the social closeness and correlation weights are set to small values (10).

Regarding the impact of the splitting strategy on the coverage of *CORALS* recommendations, Table X shows that the results are very similar across different strategies and weight values. This implies that this function is very insensitive to these factors when it comes to coverage. As future work, we intend to investigate other possible factors that might impact coverage of *CORALS*.

Finally, we perform some experiments to analyze the diversity, total of recommendations and coverage of *Affin* for a fixing level of the splitting strategy (90% - 10%) and varying the social closeness and affiliation weights, which, according to the results of the 2^k design, are more important factors. Tables XI - XII show the results for each response variable. Hence, for each factor, there is a level that provides better results for the evaluation metrics. Considering the current analysis, it is possible to properly setting the parameters of the recommendation functions *Affin* and *CORALS*.

6. CONCLUSIONS AND FUTURE WORK

We have evaluated the factors that impact the effectiveness of collaboration recommendation functions using a 2^k factorial design, where the splitting strategy and the functions' weights are the factors. Our main contribution is discovering how these factors and their interactions affect the academic collaboration recommendation regarding recall, diversity, novelty, coverage, total of recommendations and number of correct recommendations. The results show that *CORALS* is more sensitive to factors than *Affin*, mainly to the splitting strategy. This evaluation allows us to focus in the parameters that influence the recommendation results the most. After quantifying the impact of each factor and factor interaction, we have also performed experiments for each response variable impacted by the splitting strategy. The results showed which level of the splitting strategy is better for each response variable.

Table XI: Diversity: Split=90%-10%

Affin \ *Sc*	**10**	**100**
10	388.21	619.43
100	25.95	16.02

Table XII: Coverage: Split=90%-10%

Affin \ *Sc*	**10**	**100**
10	0.35	0.51
100	0.32	0.35

Table XIII: Total: Split=90%-10%

Affin \ *Sc*	**10**	**100**
10	535	6984
100	468	535

Overall, we found that all responsible variables, but coverage of *CORALS*, are significantly affected by the choice of the splitting strategy. As future work, we plan to study other ways to set the lower and upper limits of the factor levels for the 2^k factorial design. We also plan to perform experimental evaluations of the recommendation functions considering others datasets.

REFERENCES

Adams, J. Collaborations: The rise of research networks. *Nature* 490 (7420): 335–336, 2012.

Brandão, M. A. *Using link semantics to recommend collaborations in academic social networks.* M.S. thesis, UFMG, 2013.

Brandão, M. A., Moro, M. M., and Almeida, J. M. Análise de fatores impactantes na recomendação de colaborações acadêmicas utilizando projeto fatorial. In *Proceedings of Brazilian Symposium on Databases.* Recife, Brazil, 2013a.

Brandão, M. A., Moro, M. M., Lopes, G. R., and de Oliveira, J. P. M. Using link semantics to recommend collaborations in academic social networks. In *Proceedings of International Conference on World Wide Web Workshops.* Rio de Janeiro, Brazil, pp. 833–840, 2013b.

Brumec, S. and Vrček, N. Cost effectiveness of commercial computing clouds. *Information Systems* 38 (4): 495–508, 2013.

Buragohain, M. and Mahanta, C. A novel approach for anfis modelling based on full factorial design. *Applied Soft Computing* 8 (1): 609–625, 2008.

Burt, R. S. Structural Holes and Good Ideas. *The American Journal of Sociology* 110 (2): 349–399, 2004.

Cohen, S. and Cohen-Tzemach, N. Implementing link-prediction for social networks in a database system. In *Proceedings of ACM SIGMOD Workshop on Databases and Social Networks.* New York, USA, pp. 37–42, 2013.

Fouss, F. and Saerens, M. Evaluating performance of recommender systems: An experimental comparison. In *Proceedings of IEEE/WIC/ACM International Web Intelligence and Intelligent Agent Technology.* Sydney, Australia, pp. 735–738, 2008.

Freyne, J., Berkovsky, S., Daly, E. M., and Geyer, W. Social networking feeds: recommending items of interest. In *Proceedings of ACM Conference on Recommender Systems.* New York, USA, pp. 277–280, 2010.

Guimarães, S., Ribeiro, M. T., Assunção, R., and Meira Jr., W. A holistic hybrid algorithm for user recommendation on twitter. *Journal of Information and Data Management* 4 (3): 341–356, 2013.

He, J. and Chu, W. W. A social network-based recommender system (SNRS). *Data Mining for Social Network Data* vol. 12, pp. 47–74, 2010.

Jain, R. *The Art of Computer Systems Performance Analysis: techniques for experimental design, measurement, simulation, and modeling.* Wiley, 1991.

Kuo, T.-T., Yan, R., Huang, Y.-Y., Kung, P.-H., and Lin, S.-D. Unsupervised link prediction using aggregative statistics on heterogeneous social networks. In *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* New York, USA, pp. 775–783, 2013.

Lak, A., Ludwig, F., Scholtyssek, J., Dieckhoff, J., Fiege, K., and Schilling, M. Size distribution and magnetization optimization of single-core iron oxide nanoparticles by exploiting design of experiment methodology. *Proceedings of IEEE Transactions on Magnetics* 49 (1): 201–207, 2013.

Lima, E. d., Pappa, G., Almeida, J. d., Gonçalves, M., and Meira, W. Tuning genetic programming parameters with factorial designs. In *Proceedings of IEEE Congress on Evolutionary Computation.* Shanghai, China, pp. 1–8, 2010.

Lopes, G. R., Moro, M. M., da Silva, R., Barbosa, E. M., and de Oliveira, J. P. M. Ranking strategy for graduate programs evaluation. In *Proceedings of International Conference on Information Technology and Application.* Sydney, Australia, 2011.

Lopes, G. R., Moro, M. M., Wives, L. K., and de Oliveira, J. P. M. Collaboration recommendation on academic social networks. In *Proceedings of Advances in Conceptual Modeling - Applications and Challenges.* Vancouver, Canada, pp. 190–199, 2010.

Rodgers, J. L. and Nicewander, A. W. Thirteen ways to look at the correlation coefficient. *The American Statistician* 42 (1): 59–66, 1988.

Shani, G. and Gunawardana, A. Evaluating recommendation systems. In *Recommender Systems Handbook.* pp. 257–297, 2011.