Title: Unbiased one dimensional University Ranking – application based preference ordering

Authors:

András TELCS
Professor of Quantitative Methods, University of Pannonia and Assoc. Prof. of Budapest University of Technology and Economics, Department of Computer Science and Information Theory
University of Pannonia, Faculty of Economics, Department of Quantitative Methods, H-8200 Veszprém, Egyetem street 10., Hungary

Zsolt Tibor KOSZTYÁN, corresponding author
associate professor
University of Pannonia, Faculty of Economics, Department of Quantitative Methods, H-8200 Veszprém, Egyetem street 10., Hungary.
kzst@vision.vein.hu
Tel: +36 88 624 645, Mobile: +36 20 208 58 40
Fax: +36 88 624 896

Ádám TÖRÖK
Professor of Economics, University of Pannonia (UP) and Budapest University of Technology and Economics. Head, HAS (Hungarian Academy of Sciences) – UP Joint Research Unit on Regional Innovation and Development Studies.
University of Pannonia, Faculty of Economics, Department of Economics, H-8200 Veszprém, Egyetem street 10., Hungary

Unbiased one dimensional University Ranking – application based preference ordering

A. Telcs, Zs. T. Kosztyán, Á. Török

Abstract

University ranking became a global, multiplayer and multi-faceted competitive game. Players include universities, national and international ranking bodies, publishing houses, governments and other policy makers as well as students. Both competition and competitiveness analysis have several dimensions in the case of university ranking.

This is why it might be assumed the choice of ranking method or technology by the rankers is considerably influenced by their perceptions as to which kind of ranking technique could help them to achieve the maximum level of worldwide reputation. Since no adequate measure of competitiveness of university rankers is known from literature, reputation can be understood as a proxy of it.

In the following, we limit ourselves to the methodology of constructing university ranking lists, while we believe that the analysis of competition between rankers is also a challenging task for researchers. Our main focus is to produce such a ranking technique which could be regarded as a first step towards overcoming the shocking diversity of methods of university ranking.

The present paper is going to take an analytical standpoint. It attempts to offer some ranking methods that produce such unbiased, one dimensional preference lists of universities which are based solely on the partial ranks generated by

applications of students. Our ranking exercise is limited to the higher educational institutions of a country with 10 million inhabitants and over 30 universities. However, it covers ten years of application data with more than hundred thousand applications per annum.

Keywords: *university ranking, preference ordering, incomplete pairwise comparison, genetic algorithms, rank aggregation*

1. Introduction and background of the study

University ranking[1] is, in the first place, about competition between higher education institutions. Furthermore, it is the product of a service industry with institutions preparing university ranking lists as players. Therefore, this is a very special and complex case of competition and competitiveness analysis. Competition taking place between ranking institutions ("rankers") is not only influenced by the quality of their respective ranking methods and ranking lists, but also by the acceptance of their ranking lists by all the interested parties (the latter include students and their families, university employees, current and future employers, government officials and many other players in politics and the economy).

As in many other competitive situations there is no such thing as perfect ranking method and/or ranking in this case either. Indicators, weights, and methodologies used all have impact on the final result of ranking while there is no one universal benchmark to compare to and judge on the relevance and correctness of the ranking[2]. It is also influenced by preconceptions, common beliefs and certain interests of players.

The technology of university ranking reached a high level of sophistication in recent years. This special field of quantitative economic analysis can be regarded from different angles. It may be expected to produce a global score in a very particular service sector giving thus a certain picture of competition. It is also an applied statistics and econometric exercise. Last but not least, it is a point of

[1] Universities are only a subset, albeit the by far most important one of the set of higher education institutions. Literature usually speaks of „university ranking lists" which is, in a strict sense of the term, inaccurate. The reason is it does not cover other higher education institutions such as colleges or some high-level schools with the intellectual capacity of a university but without such a name (e. g. the London School of Economics. To be fair however, we must note that 1. it is very rare that colleges figure on international higher education ranking lists; and 2. special schools enjoying university reputation are usually considered as such also by the authors of ranking lists. This is why we, although with some reluctance, also refer to „university ranking lists" in this study.
[2] On this diversity, see for example: Török (2009); Shin et al. (2011).

reference for governments and policy makers as well as a lucrative business for publishing houses.

Enormous literature is available on the history, development and recent practice of university ranking (as a starting reference see Vught, Ziegel 2012), and Shin (2011) Harman (2011), Teicher (2011), and references there, also see Török (2009)). An international effort initiated by the European Commission has been undertaken to develop a multidimensional hence more flexible ranking system of universities in Europe. That report Vught, Ziegel (2012) concludes that accuracy, relevance and availability of data constitute the key challenges for any indicator-based ranking of higher education institutions. Other issues widely debated among researchers and practitioners include indicator weighting and the requirements of a rigorous interpretation of the output of the analysis.

Having seen the complexity of the problem we may start with the hypothesis that there is no such thing as a single perfect and waterproof ranking method of higher education institutions both on the national and the international level.

The present paper is the first one in a series in which we try to investigate the competitive positions of higher education institutions (universities in short) trying to pinpoint the pitfalls rankers may face. In the present paper we develop an unbiased, one dimensional preference list of universities based solely on the partial ranks generated from student applications. At this stage of our research, we compare the ranking we obtained with results of different sources. In later works we plan to incorporate university indicators, local, regional and economic indicators to analyze the choices of students.

In general the problem to create the best fitting (lowest cost) linear order, based on a complete or incomplete weighted directed graph, is very difficult. Several heuristic methods have been developed to obtain an approximate solution e.g. Martí, Reinelt, (2011).

In what follows we focus on the construction of the aggregate of the applicants' preference list. Our data cover the 2001-2010 period annual university applications in Hungary.

2. The methods

2.1. The source of data

The rough data source is the Hungarian national center of higher education (HE) applications - Educatio Nonprofit Ltd. We shall refer to the agency as APPLI mirroring the common reference to it in Hungarian. APPLI collects and handles all HE applications. Their database contains the annual applications. Our subset contains 10 years of application records, and each has ten fields.

1. Year
2. Semester
3. Student ID
4. Number of personal preference order
5. HE institution
6. Faculty/School
7. Course
8. Level of study
9. Form of study
10. Government or private financing of tuition.

Each record refers to a single application. One student may make more than one. It is typical that a student applies to three places but there are cases of more than ten applications to different higher education institutions. Our database contains more than 400 000 records per annum from more than 100 000 applicants. From the point of view of an applicant it means as many records as applications he or she has ordered in the field #4 according to her or his preference.

It is clear from this structure that we may lose a lot of information and the outcome will be very biased if only the HE institution is considered without respect to the Faculty or School (e. g. Law, Engineering, Medical or other) within the institution where courses to be taken by the applicant are offered. Imagine that a student named the course C1 at faculty F1 of university U1 as his/her first preference, a completely different C, F, U as second and a course C2 from faculty F1 of university U1 as third. How can we interpret that information in a final ranking? Of course at the very end such a mix will be inevitable, but we decided to store the input data without any loss of information.

We will filter the dataset later with respect to the form and financing of the studies in order to reduce the number of such ambiguous cases. Also we shall consider well-defined fields of studies and create preference lists of courses

instead of entire universities at least in the phase of thorough evaluation of our proposed methods. Therefore, we present a university ranking list only at the final stage of this work.

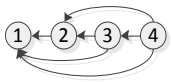
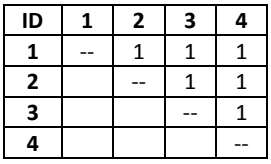
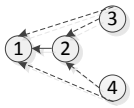
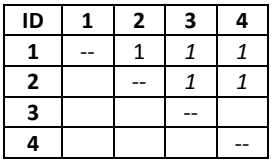
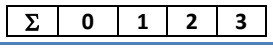
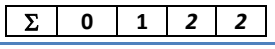
2.2. The data representation

To ease the terminology in the sequel we shall use the term university instead of object (to be ranked and might be university, faculty or course until it is not specified otherwise.) Let m be the number of courses and n the number of applicants. The i^{th} applicant choice vector is $\mathbf{a}^i:=[a^i_1,\ldots,a^i_{m_i}]^{\mathbf{T}}$, $i:=1,2,\ldots,n$. of length m_i less than m or equal to m, the set of vectors is $\mathbf{A}:=\{\mathbf{a}^1,\ldots,\mathbf{a}^n\}$.

The individual choices are coded in the matrix of objects (courses or later faculties, universities). The matrix is the adjacent matrix of an oriented graph on the vertices, representing objects. If applicant i preferred object k to l then there is an oriented edge from l to k. Now some cautionary notes are in order.

1. The oriented edges are multiple edges in the final graph.
2. If i has preference list [1,3,2,4] then edge points to 1 are not only from 2 but from 3 and 4 as well.
3. The unranked universities are less preferred than any of the named ones.

Table 1: Graph representation of the application

Individual preference graph	Adjacency matrix	Individual preference graph	Adjacency matrix
$\mathbf{a}^1:=[1,2,3,4]^T$, $m=4$. 		$\mathbf{a}^2:=[1,2]^T$, $m=4$. 	
Sum:		Sum:	

Let us observe that in the second example the unranked universities (3,4) have the same role hence it is rational to treat them in the same way. University 3 an 4 have not been chosen, both are less preferred than 1 or 2 and meanwhile there is no preference order between 3 and 4. The graph representation is intend to reflect this situation and orient edges from 3 (and 4) toward 1 and 2 and at the same time no edge is defined between 3 and 4. Later an other equivalent representation will be given, in which two edges will be defined one from 3 to 4

and an other oppositely and half weights will be assigned to them. We shall see that the latter representation is also suitable for the ranking problem.

2.3. The cost function

Once we have a preference list of the universities we have to judge its correctness compared to the individual partial, preference lists. The comparison should be based on a penalty or cost function *h*. Its definition is crucial in the evaluation of the ranking methods. The Kemeny-Young method (see Kemeny 1959) is widely accepted to measure the correctness of a ranking. In that we consider the given full ranking as an oriented path and count all the oriented edges pointing in the opposite direction. This counting can be represented in the matrix scheme easily.

Let **M** be the (*m* by *m*) adjacency matrix of the oriented graph of applications. The element m_{ij} $i \neq j$ shows how many times the university *i* was preferred against *j*.

2.3.1. Reverse order penalty

Now let $\mathbf{b}=[b_1,b_2,\ldots,b_m]^T$ be an arbitrary ordering, furthermore let $\mathbf{M_b}$, be the column rearrangement of the matrix **M** so that the columns follow the order of **b.** The cost or penalty function can then be defined as follows:

$$h(\mathbf{M},\mathbf{b}) = \sum_{i=1}^{m}\sum_{j=i+1}^{m} m_{\mathbf{b}ij},\ \forall m_{\mathbf{b}ij} \in \mathbf{M_b}. \qquad (1)$$

That means exactly how many times were preferences in the opposite order than in **b.** One can see that the cost function is nothing else than the sum of the elements below the diagonal of $\mathbf{M_b}$.

2.3.2. Least squares

There are other possibilities to measure correctness of a ranking **b**. The usual square error is applicable

$$\sum_{i=1}^{n}\sum_{j=1}^{m}\left(a_j^i - b_j\right)^2, \text{ where } a_j^i \in \mathbf{a}^i, b_j \in \mathbf{b}. \qquad (2)$$

It is known that the Borda-Kendal method minimises (2) Kendal (1962)

2.3.3. Measure of compliance

The pairwise comparison method involves another measure of compliance, the stress.

$$stress: = \sqrt{\frac{\sum_{i=1}^{m}\sum_{j=1,i\neq j}^{m}\left(pcm_{i,j}-p_{i,j}\right)^2}{\sum_{i=1}^{m}\sum_{j=1,i\neq j}^{m}pcm_{i,j}^2}} \qquad (3).$$

This measure compares the aggregate pairwise comparison matrix to the one given by the ordering again in square errors.

Later we shall compare methods using those cost functions without elaborating on the difference of their inherent nature.

2.4. Node degree ranking

The simplest rank aggregation is based on the out-degrees of the directed graph. This method assumes that a university has lower position in the ranking if it has higher out-degree, i.e. the more universities are preferred relative to it the lower the preference is. Ranking is then quick and easy. Consider the preference matrix **M** and calculate the column sums. The increasing order of the sums is the degree rank **b**. Rearrange the columns according to **b**. The method is demonstrated in Table 2.

Table 2: The node degree ranking

The adjacency matrix (M)

ID	1	2	3	4
1	--	5	5	3
2	3	--	1	5
3	2	3	--	2
4	1	5	3	--
Σ	6	13	9	10

$h(\mathbf{M},[1,2,3,4]^T)=17$

$\mathbf{b}=[1,3,4,2]^T$

The rearranged adjacency matrix (M_b)

ID	1	3	4	2
1	--	5	3	5
3	2	--	2	3
4	1	3	--	5
2	3	1	5	--
Σ	6	9	10	13

$h(\mathbf{M},[1,3,4,2]^T)=15$

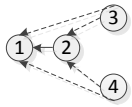
It should be noted that the method may lead to ties - the final order is not unique - since column sums may coincide. In case of high number of edges the coincidence of column sums is very unlikely. In such a case the cost function has no global minimum.

2.5. The rank sum method

The rank sum method is widely used and known. Here we can demonstrate it using the graph representation of preferences in a very concise way. This method differs from the node degree method only in the handling of the unranked universities. In this model the unranked entries are pointing to the ranked ones with an oriented edge but the edge has a weight different from one. In other words the individual incomplete preference vector $^{i}=[a^{i}_{1},\dots a^{i}_{k}]$ ($k=k_i$ the number of ranked objects) is completed with the average rank of unused ranks. That is, all unranked universities got the same rank $\boldsymbol{r}=\frac{\sum_{l=k_i+1}^{k} l}{m-k_i}$. The i^{th} student has the rank vector $\mathbf{s}^i \in R^m$, where $\mathbf{s}^i := \left[s_1^i, s_2^i, \dots, s_m^i\right]^T$, and

$$s_j^i = \begin{cases} p, \text{ if } a_p^i = j \in \mathbf{a}^i, \\ \frac{1}{m-k_i}\sum_{l=k_i+1}^{m} l \end{cases} \qquad (4)$$

Table 3: edge weights for the rank sum method

Subgraph of individual 1's preferences	Matrix of preferences	New weights ($l>m_i$)
		$x=\frac{\tilde{s}_l^i}{k_i}=\frac{s_l^i-1}{k_i}\frac{\sum_{l=k_i+1}^{m} l}{m-m_i}-1=\frac{\sum_{l=k_i}^{m-1} l}{k_i(m-k_i)}$, $l\notin\mathbf{a}^i$

ID	1	2	3	4
1	--	1	x	x
2		--	x	x
3			--	
4				--
Σ	0	1	2x	2x

Sum:

Corrected preference vector: $\tilde{s}^i$ 0 1 2.5 2.5

x=1.25.

The rank-sum and the edge weight sum should be equal, that needs normalization and yields that the edge weight defined as

$$p := \begin{cases} 1, & \text{if} \quad a_g^i = j, a_h^i = p \in \mathbf{a}^i, g < h \\ \frac{\sum_{l=k_i}^{m-1} l}{k_i(m-k_i)}, & \text{if} \quad j \in \mathbf{a}^i, k \notin \mathbf{a}^i \\ 0, \text{otherwise.} \end{cases} \qquad (5)$$

It is worth to note that the rank-sum and node degree method differs only in the handling of unranked universities (or more generally the ties, as discussed below Table 1.).

2.6. Pairwise comparison method for partial ordered lists

The pairwise comparison method is rooted in the thorough investigation method of survey taking in which respondents compare all the pairs of objects under scrutiny. This method is free of information loss and provides a very accurate picture on respondent opinions. In our case the students do not provide pairwise comparisons but a linear preference order. Nevertheless the pairwise preference can be read out from this strict ordering without ambiguity. Since we have a partial list we have to turn to the recent extension of the pairwise comparison method to ones based on incomplete comparisons. As previously we assume that the unranked universities are less preferred to all the designated ones. We also assume that the preference is neutral between any unranked universities. Several solutions have been published Alonso & all (2005), Tannino (1984) and Farkas (2003). For the incomplete comparison problem, we adopt the method of Fedrizzi and Giove (2007).

The PCM (Pairwise Comparison Matrix) input matrix is obtained from the aggregated preference matrix **(M)**. Let the element of the PCM, for all i,j

$$pcm_{i,j} = \begin{cases} \frac{m_{i,j}}{m_{i,j}+m_{j,i}}, \text{if } m_{i,j} + m_{j,i} > 0 \\ 0{,}5 \text{ otherwise.} \end{cases} . \qquad (6)$$

The diagonal elements of the PCM are neglected. The sum of symmetric pairs, $p_{i,j}+ p_{j,i}$ equals to one. The value $p_{i,j}$ reflects the proportion of students preferring i to j.

The method then can be carried out in a nutshell as follows. A new matrix **Z** is created. The matrix entries represent the difference of row and column differences of **PCM**. It can be seen that the row and column sums of **Z** are zero. Let z_i denote the column sum of **Z**. As a result we have a non-negative real z_i, for all entities. The increasing order of z_i -s provides **b** as an estimate of the preference list of entities.

There is an elegant verification of the correctness of the final scores. We introduce **D** the distance matrix of the preference values $d_{i,j}=z_i-z_j$, $i\neq j$ and then calculate a matrix **P** within which $p_{i,j}$, the entries correspond to the p-values of $d_{i,j}$ of the standard normal distribution. The comparison of the matrices **P** and **PCM** done with the well-known χ^2 test. The test function is called in this context *stress* value and expressed by

$$'ess: = \sqrt{\frac{\sum_{i=1}^{m}\sum_{j=1,i\neq j}^{m}\left(pcm_{i,j}-p_{i,j}\right)^2}{\sum_{i=1}^{m}\sum_{j=1,i\neq j}^{m}pcm_{i,j}^2}} \qquad (7).$$

Small stress values indicate that the estimate of the preference list is correct.

2.7. Genetic algorithms

Genetic algorithms (see Mitchell, M. 1996) are widely used solving NP hard problems (as the linear ranking is being equivalent with the travelling salesman problem) and other complex optimization task. GA-s are successfully applied where a good approximation of the optimum is acceptable while to get the perfect optimum is impossible or very difficult due to the lack of closed form solution or due to the fact that the computation is beyond reach.

We are facing a version of the travelling salesman problem for which several GAs are developed in Braun (1991). In our case the fitness function is the inverse cost function, the competing species are the permutation of the objects and the genes are the positions of the objects in the permutations.

If **b** is a list of entities as above the cost function $h(\mathbf{M},\mathbf{b})$ can be calculated easily. Simply the columns of M are rearranged according to the increasing order in **b** and summing up the cells below the diagonal. The graph interpretation of this is the following. The nodes are the entries in **b** and $h(\mathbf{M},\mathbf{b})$ is the cost of the Hamilton path along **b**. The task is to find the minimal cost Hamilton path, which is a travelling salesman problem on the full graph. The genetic algorithm is not using the edge costs but the cost function.

In our GA setup a permutation of the entity is a specimen in the population. The fitness function of a specimen is the inverse cost function. As usual two operations modify the genes, recombination and mutation. Let us illustrate them with very simple examples. Let m=10 and consider a simple transposition in the ordering $[1\ 2\ 3\ \mathit{4} \leftrightarrow \mathit{10}\ 5\ 6\ 7\ 8\ 9]^T \rightarrow [1\ 2\ 3\ \mathit{10}\ \mathit{4}\ 5\ 6\ 7\ 8\ 9]^T$), that is an elementary mutation. The recombination is based on two sequences of genes. The input sequences are [*1 2 3* 4 6 10 5 7 8 9] and [3 2 1 *6 4 10 9 8 7 5*] their recombination is [*1 2 3 6 4 10 9 8 7 5*]. In the parameterization of the GA we followed the method presented in the paper (Braun 1991).

2.8. Data aggregation

In our rough dataset entries are courses provided by faculties of universities. In our investigation we might be interested in the ranking of faculties or

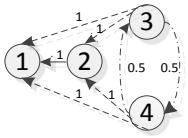
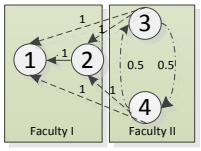
universities not only competing courses. The picture is quite complex; there are cases when the same course is offered not only by different universities but also by different faculties at the same university. The graph representation of preference lists provides an easy and transparent way to obtain higher-level rankings. Data aggregation means in the graph aggregation of nodes and edges receive sum of the edges connecting the aggregated nodes.

2.8.1. The aggregation in details

All the courses receive a unique identifier $i=1,\ldots,n$, those are the nodes of the graph. On an aggregate level, a faculty, university or region has its own identifier, k,l,m and each object below this inherits this identifier. For instance a course carries the identifier of the faculty, of the university and of the region as well. If the aggregation is on the level of faculties we have a graph of faculties (nodes with the faculty identifiers) and all the edges between the courses of the same faculty are neglected, meanwhile the edges between courses of different faculties kept and form multiple edges between the faculty nodes.

Let us consider an example. We have m=4 courses belonging to faculty I and II. The course records are (1,I) and (2,I) and (3,II) and (4,II) . A student preference list was $\mathbf{a}^i:=[1,2]^T$.

Table 4: node aggregation

	Preferences between courses	Aggregation to faculty level	Preferences between faculties
Preference graph			

Preference matrix

ID	1	2	3	4
1	--	1	*1*	*1*
2		--	*1*	*1*
3			--	0.5
4			0.5	--

ID	1	2	3	4
1	--	1	*1*	*1*
2		--	*1*	*1*
3			--	0.5
4			0.5	--

ID	I	II
I	--	4
II		--

In our example the student preferred 1 to 2 while 3 and 4 are not ranked, in our convention less preferred as 1 and 2 as well. As a result in the aggregate graph we have one link from II to I with weight 4. That is reflected by the matrix of faculties in Table 4. In other words courses of the same faculty merged into one node of the faculty-graph and the oriented edges are also merged into edges between of the higher level nodes, faculties in our example. In the same way

faculties of the same university can be merged and edges inherited, merged into edges between the universities.

The aggregation procedure works in the same way between any two levels.

2.9. Test of the methods

The correctness and efficiency of the selected methods is tested on synthetic data.

The tests work as follows. There is a given order (w.l.g $\mathbf{b}=[1,2,\ldots n]^T$), we call it target) to be reconstructed from a set of partial orders. The data set contains perturbation of the target. The robustness and efficiency investigated against different perturbations. The simplest one is if the target is modified uniformly and randomly and the resulted sample set is censored according a given upper percentile and Euclidian distance from the target.

2.9.1. Uniform sample

Now we describe this perturbation and test it in detail. We generate random preference vectors of the same length as the target for seek of simplicity. (It might be appropriate to generate partial preference list of much smaller length and fill them up as described above but we omit this case since the second test we generate partial preference lists close to real life situation). The pseudo-code of the list generation is presented below. The censoring of the randomly generated list is done by the function

$$f(x) := \begin{cases} 0, \text{if } x < 0 \\ \lambda e^{-\lambda x}, \text{if } x \geq 0 \end{cases} \qquad (8)$$

This ensures that the closer is a permuted list is to the target the higher is the probability to be selected into the data set.

2.9.2. Choice simulation

In order to run a test on a dataset close to the real one we simulate students' choice generating a sample based on the multinomial logit model. Let us assume that the choice depends on t parameter. Denote $U_{a_j^i}$ the utility of the j-th

course based on the *t* parameter for student *i* and $\mathbf{a}^i := [a_1^i, a_2^i, \dots, a_m^i]^{\mathrm{T}}$ the ranking list of the *i*-th student, based on the order of utility values: $U_{a_1^i} < U_{a_2^i} < \cdots < U_{a_m^i}$. For seek of simplicity we assume that all universities are ranked. As usual, the utility is assumed to depend on the parameters $U_{a_j^i} = V_{a_j^i} + \varepsilon_{a_j^i} = \beta_{1,a_j^i} x_{1,a_j^i} + \beta_{2,a_j^i} x_{2,a_j^i} + \cdots + \beta_{t,a_j^i} x_{t,a_j^i} + \varepsilon_{a_j^i,}$ where $\varepsilon_{a_j^i}$ are i.i.d. logistically distributed random variables. The weights of the parameters for the *i*-th students are $\beta_{1,a_j^i}, \beta_{2,a_j^i}, \dots, \beta_{t,a_j^i}$. The probability of the final order $= [b_1, b_2, \dots, b_m]^{\mathrm{T}}$, based on the utilities is

$$\mathbf{P}\left(U_{b_1} < U_{b_2} < \cdots < U_{b_m}\right) = \prod_{j=1}^{m} \frac{\exp\left(\mathbf{V}_{b_j}\right)}{\sum_{l=j}^{m} \exp\left(\mathbf{V}_{b_l}\right)}, \tag{9}$$

where $\mathbf{V}_{b_l} = \left[V_{a_l^1}, V_{a_l^2}, \dots V_{a_l^n}\right]^{\mathrm{T}}$ the "error free" utilities. We generate sample elements according this model, we assume that the individual *weights* are normally distributed and i.i.d.. The ranking methods can be tested and compared on such data sets which model real life individual choices.

3. Examples

The ranking methods described above are tested on the application data of 2011. The courses of physics at BA level and MBA programs are chosen. The aggregation is demonstrated on the aggregated data of the applications of Bologna type education and the earlier 5 system[3]. In case of BA applications only government financed study places are included, while in the case of the MBA self financed applicants are also included. The results are summarized in the tables below.

3.1. Comparison of methods on 2011 application data

Table 5: Preference list of students applied to business informatics state financed positions in 2011 (Methods: RS=Rank Sum, CS=Column Sum, PW = Pairwise, GA= Genetic Algorithms) (For the acronyms of the universities see the Appendix)

Methods (RS, CS, PW, GA, PR)	Ranksum (RS)	Pairwise comparison (PW)	Page Rank (PR)	Application	1st ord. Application

[3] Hungary had traditionally five year long university studies in most of the cases till the adaptation of the Bologna 3+2 system.

Pref. ord.	INSTITUTE	Mean Rank	*Z*	$\Phi^{-1}(Z)$	*PR*(*E*)	2011	2011
1	SZTE	6.0901	0.0860	0.5343	0.3065	506	467
2	BCE	6.1234	0.0790	0.5315	0.3021	477	453
3	DE	6.1386	0.0758	0.5302	0.3057	483	435
4	BGF	6.2256	0.0588	0.5234	0.2762	442	286
5	SZE	6.3827	0.0247	0.5099	0.2776	362	307
6	PTE	6.4638	0.0077	0.5031	0.2778	323	270
7	PE	6.6927	-0.0402	0.4840	0.2699	207	168
8	ME	6.7193	-0.0458	0.4817	0.2791	196	153
9	KRF	6.7238	-0.0467	0.4814	0.2901	194	154
10	DF	6.7536	-0.0529	0.4789	0.2940	181	136
11	ZSKF	6.8170	-0.0663	0.4736	0.3081	150	77
12	NYME	6.8835	-0.0802	0.4680	0.2731	115	81
Error value *h*(M,b):		92818,5	**Inhomogenity index** (*I*)		**47.08%**		
stress:		0.0043					

In the business informatics BA case all the methods produced the same ranking. The first university got 506 applications, while the 12th only 115, but this difference is not reflected so markedly in the preferences. The indicators like *PR*(*E*), or *Z* for the pairwise comparison method as well as the average rank values are close to each other.

Table 6: The preference list of students applied to MBA in 2011-ben (Methods: RS=Rank Sum, CS=Column Sum, PW = Pairwise, GA= Genetic Algorithms)

Methods (RS, CS, PW, GA, PR)		**Ranksum (RS)**	**Pairwise comparison (PW)**		**Page Rank (PR)**	**Application**	**1st ord. Application**
Pref. ord.	INSTITUTE	Mean Rank	*Z*	$\Phi^{-1}(Z)$	*PR*(*E*)	2011	2011
1	PE	3,1593	0,3925	0,6527	0,5150	219	166
2	BME	3,8488	0,0822	0,5328	0,3909	102	94
3	ME	3,9140	0,0647	0,5258	0,3876	98	85
4	DE	4,2640	-0,0567	0,4774	0,3529	54	40
5	PTE	4,3977	-0,1153	0,4541	0,3329	35	29
6	BCE	4,5791	-0,1759	0,4302	0,3162	14	9
7	SZTE	4,6093	-0,1914	0,4241	0,3107	9	7
Error value *h*(M,b):		3810	**Inhomogenity Index (*I*)**		**42,19%**		
stress:		0,0038					

In the MBA example the mix of government and self-financed applications is considered. Still, the preference matrix can be reconstructed from the Z-values,

as the low stress index indicates. The average rank values except for the first one are close to each other, which is also reflected in the high homogeneity index.

Different methods resulted in the same order. The details provide additional information.

If we study the applications to 39 universities, the ordered lists show some minor differences. The next table shows the universities ranked to the first ten positions.

Table 7: The preference list of applications to the master and the traditional one stage education in 2011 (Methods: RS=Rank Sum, CS=Column Sum, PW = Pairwise, GA= Genetic Algorithms, PR=Page Rank)

Methods	CS, RS	PW	GA	PR
Pref. ord.	INSTITUTES	INSTITUTES	INSTITUTES	INSTITUTES
1	ELTE	ELTE	ELTE	ELTE
2	SZTE	SZTE	SZTE	SZTE
3	DE	DE	DE	DE
4	PTE	PTE	PTE	SZIE
5	BCE	BCE	BCE	BCE
6	BME	BME	BME	PTE
7	SE	SE	SZIE	BME
8	PPKE	SZIE	SE	SE
9	SZIE	PPKE	PPKE	PPKE
10	ME	ME	ME	ME
***h*(M,b)**	16608408.5	16607451.5	16608450.5	16607294.5
I	48.29%	48.29%	48.29%	48.29%
		Stress	0.0018	

One can see that the smaller error is produced by the genetic algorithm, the biggest among the heuristic methods produced by the Page Rank method, and the lowest by the pairwise comparison. On the first six positions all the heuristic methods coincide, and there are only small position swaps below. The Page Rank method agrees only at the first three places with the others. The best methods, the genetic algorithm and the pairwise comparison method produce only a single difference in the ordering. The rank correlation of the lists of different methods is shown on Table below.

Table 8: The rank correlation of the lists(Methods: RS=Rank Sum, CS=Column Sum, PW = Pairwise, GA= Genetic Algorithms)

Rank corr.	CS,RS	PW	PR	GA
CS,RS	1.000	0.991	0.498	0.892

PW	0.991	1.000	0.507	0.912
PR	0.498	0.507	1.000	0.507
GA	0.892	0.912	0.507	1.000

3.2. Test of methods on synthetic data

In the previous section we described the generation of random, individual preference lists. It is not intended to model students' behaviour, but some properties of the applications are taken into consideration. The majority of the applications contained maximum three courses since listing more than tree needed another form to fill in. Less than 1% of the applications contained ten or more courses, so we limited the length of the simulated applicant's preference list to 10.

First scenario. Let **b**=[1,2,..10]. The random preference lists (for which the distance from b is measured with the Euclidian distance (see e.q. 7)) distributed by $\lambda e^{-\lambda x}$, where x is the distance from **b** and λ is the parameter of the distribution set to λ_1=1; λ_2=0.1; λ_3=0.01.

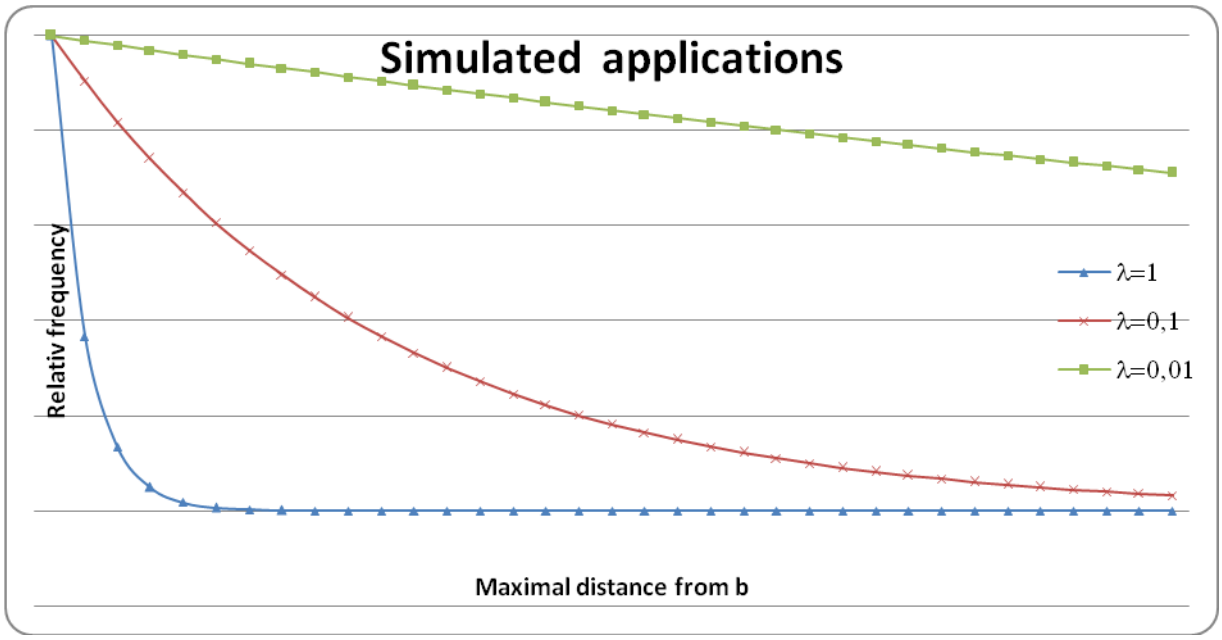


Figure 1: Rankings of simulated application

Results are summarized in Table 9 below.

Table 9: Results of simulated applications

OBJID	λ=1	λ=0.1	λ=0.01

	Mean Rank	Z	$\Phi^{-1}(Z)$	$PR(E)$	Mean Rank	Z	$\Phi^{-1}(Z)$	$PR(E)$	Mean Rank	Z	$\Phi^{-1}(Z)$	$PR(E)$
1	1.0420	1.5280	0.9367	0.9593	2.4027	1.0714	0.8580	0.5274	4.1500	0.3668	0.6431	0.3971
2	1.9683	1.1255	0.8698	0.2780	2.5293	1.0582	0.8550	0.5136	4.1947	0.3526	0.6378	0.3945
3	2.9914	0.7320	0.7679	0.0489	3.0530	0.8622	0.8057	0.4412	4.4722	0.2956	0.6162	0.3717
4	3.9983	0.3765	0.6467	0.0086	3.9689	0.5401	0.7054	0.3460	5.0917	0.1334	0.5531	0.3264
5	5.0002	0.0595	0.5237	0.0016	5.4094	-0.0142	0.4943	0.2303	5.2027	0.0818	0.5326	0.3262
6	6.0028	-0.4560	0.3242	0.0004	6.3460	-0.3093	0.3785	0.1827	5.3167	0.0478	0.5190	0.3238
7	7.0125	-0.8058	0.2102	0.0001	7.0188	-0.5091	0.3054	0.1506	5.7594	-0.0656	0.4739	0.2888
8	8.0418	-1.0965	0.1364	0.0000	7.3132	-0.5594	0.2879	0.1475	6.0257	-0.1331	0.4470	0.2735
9	8.9428	-1.4631	0.0717	0.0000	8.3234	-0.9567	0.1694	0.0894	6.7814	-0.3396	0.3671	0.2282
10	9.9580	-2.0231	0.0215	0.0000	8.6353	-1.1833	0.1183	0.0699	8.0055	-0.7397	0.2298	0.1432
	I	**0.25%**	***stress***	**0.0055**	***I***	**16.11%**	***stress***	**0.0135**	***I***	**45.41%**	***stress***	**0.0234**

It is easy to recognize that the smaller the λ the bigger the inhomogeneity and the stress, while all the methods reconstruct the original order **b**.

3.3. The simulated logit

The next simulation has been designed to test the robustness of the methods on logistic data. Data are generated by a logit model based on synthetic parameters, not related to real values of university utilities. Meanwhile we keep in mind that in a forthcoming study the logit model and our ranking methods will be used and results matched. We assume that the utility is determined by three factors (one may picture that those are the distance between home and university, the others are compulsory fees and faculty credits). The contribution to the overall utility is positive for one and negative for two factors, all having a unit absolute value. The unknown individual β coefficients are chosen randomly. The generated utilities are summarized in Table 10 below.

Table 10: Simulated coefficients and the utility values

	x.	β_1.	β_2.	β_3.	β_4.	β_5.	β_6.	β_7.	β_8.	β_9.	β_{10}.
x_1	1	-0.05	-0.06	-0.06	-0.16	-0.17	-0.58	-0.55	-0.71	-0.81	-1.00
x_2	1	-0.25	-0.28	-0.17	-0.21	-0.13	-0.27	-0.40	-0.40	-0.29	-0.32
x_3	1	0.18	0.15	0.02	0.10	0.00	0.03	0.04	0.14	0.09	0.16
V.		-0.12	-0.20	-0.22	-0.26	-0.30	-0.82	-0.91	-0.97	-1.00	-1.16
rank		**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**

In order to have the same number as the number of applications in 2011 in Hungary, we generated 161731 random utility values so that $U_{ij}=V_i+\varepsilon_{ij}$ (i=1..10; j=1..161731), where ε_{ij} follow logistic distribution with m=0 and b=1

parameters. The original order of the objects was $\mathbf{b}=[1,2,..,10]^T$ and the different methods were tested on the randomly generated utilities U_{ij}.

Table 11: Results of the different methods on full simulated lists

Pref. ord.	Mean Rank	*Z*	$\Phi^{-1}(Z)$	*PR*(*E*)
1	3.2543	0.6681	0.7480	0.4517
2	5.0303	0.5406	0.7056	0.4400
3	5.1982	0.1622	0.5644	0.3405
4	4.9706	0.1438	0.5572	0.3348
5	3.5575	0.1234	0.5491	0.3254
6	8.4251	0.1032	0.5411	0.3060
7	6.3743	0.0803	0.5320	0.2937
8	5.3050	-0.2337	0.4076	0.2285
9	5.1297	-0.7008	0.2417	0.1449
10	7.7845	-0.8869	0.1876	0.1238
***stress*=**	0.0147		*I*=	30.22%

The results listed in Table 11 above are based on the data simulating where all applications contain all the universities in our sample thus providing a full ranking. We can refine this picture taking into consideration the distribution of the length of the preference lists, created by the applicants. The frequency of the length of the partial ranking list is shown in the second column of Table 12. The simulated dataset is prepared, full lists are truncated so that it contains the same number of partial preference lists of given length as the real applications. All the above discussed methods are tested on this data set.

Table 12: Results on simulated partial lists (Methods: RS=Rank Sum, CS=Column Sum, PW = Pairwise, GA= Genetic Algorithms, PR=Page Rank)

Number of appl.		Orig. ord.	RS=CS	PW	GE	PR
1	161,731	1	1	1	1	1
2	128,286	2	2	2	2	2
3	107,070	3	3	3	3	4
4	42,738	4	4	4	4	3
5	26,008	5	5	5	5	6
6	16,906	6	7	6	6	5
7	7,365	7	6	7	7	8
8	5,132	8	9	8	8	7

9	2,995	9	8	10	9	9
10	2,215	10	10	9	10	10
Computational Demand (Pentium core 2 duo, 4GB RAM):			120 ms	210 ms	2,782 ms	86 ms
		***stress*:**	0.0459	***I***	33.85%	

One can find that the Page Rank method is the quickest but it does not reconstruct the original order. The genetic algorithm is the slowest but able to reproduce the original order. The other methods provide scores as well, GA not. The pairwise comparison method provides the mutual distances as well.

4. Final remarks

The proposed graph representation of data is able to reflect all the information coded in the applications without loss of data.

The methods investigated properly aggregate partial preference lists and provide a single "optimal", best fitting one. The problem to find the optimal linear order is an NP-hard combinatorial task, only heuristic solutions are feasible. If we want to establish the institutional ranking the rank sum methods seems to be the best, if the relative positions are also needed, we propose to use the pairwise comparison method.

4.1. Future research

Based on the results presented we can use methods to develop preference lists based entirely on students' choices. On the other hand several university rankings have been published, all of which are based on complex methodologies. Those take into consideration several characteristics of universities and use same weighted combination of them in order to create rankings. Other methodologies provide multidimensional rankings. Those contrary to the name still based on weighted combination of several characteristics of the universities, but those characteristics are grouped into different dimensions like research, quality of education etc. We plan to follow a totally different approach. Using any of the methods proposed and tested above

we generate a linear preference list solely based on students' choice and try to understand their perception of university characteristics via the preference lists.

We would like to demonstrate our approach with one final example. A Hungarian business weekly called HVG publishes a university ranking each year, which is basically identical to the online ranking http://eduline.hu/rangsor. This ranking is based on two major dimensions. One is students' achievements (collected entry criteria score, number of language certificates, student contest credits), the other is faculty excellence (research performance) and their per student ratio.

Table 13 below shows the top 12 institutions (from http://eduline.hu/rangsor) ranked by the institute excellence and students' preference developed by the genetics algorithm. The rank correlation of the two 2011 lists is 0.329.

Table 13: Faculty excellence and student preference list calculated by GA method (from 2001-2011)

Rank	Student preference list (GA)											Faculty Excellence
	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2011
1	ELTE	ELTE	ELTE	ELTE	ELTE	ELTE	ELTE	ELTE	ELTE	ELTE	ELTE	ELTE
2	SZTE	DE	BGF	SZTE	SZTE	SZTE	SZTE	SZTE	DE	DE	DE	SZTE
3	PTE	SZIE	SZTE	DE	PTE	PTE	DE	DE	SZTE	SZTE	SZTE	DE
3	DE	SZTE	PTE	PTE	DE	DE	PTE	PTE	PTE	PTE	PTE	NYME
5	BGF	PTE	DE	BGF	BCE	BCE	BCE	BCE	BCE	BCE	BCE	SE
6	SZIE	BGF	SZIE	BKÁE	BGF	BGF	BGF	BME	BME	BME	BGF	PTE
7	NYF	NYF	BKÁE	NYF	BME	BME	BME	BGF	BGF	SZIE	SZIE	KRE
8	KJF	BKÁE	BME	SZIE	NYF	NYF	ME	NYME	SZIE	BGF	BME	BCE
9	BME	BME	NYF	SZE	PPKE	ME	BMF	ME	ME	NYME	NYME	ME
10	ME	ME	ME	BME	VE	SZIE	NYF	SZIE	NYME	ME	ME	PE
11	EKF	KJF	BMF	ME	SE	PPKE	SZIE	PE	BMF	SE	SE	BME
12	TSF	TSF	EKF	BMF	EKF	EKF	SE	BMF	PPKE	BMF	SZE	NYF

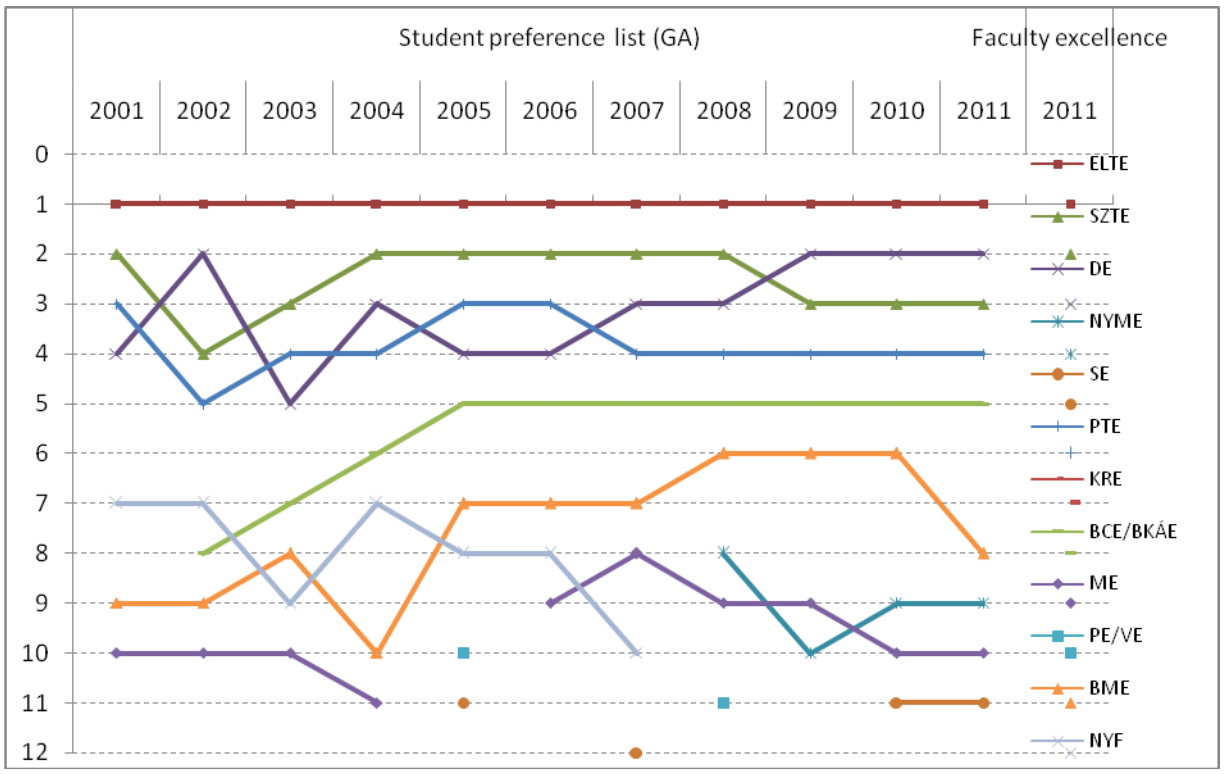


Figure 2: Faculty excellence and student preference list calculated by GA method (from 2001-2011)

The first position is met, the next two are flipped, the others show a very mixed picture. We know that the different aspects of faculty excellence are measured accurately, but their combination weights are artificial (subjective). Our ordering method is completely free of such bias, hence provide the opportunity to find the real factors, weights students assign to different aspects of faculty performance. Our work is not going to replace university rankings. We do not intend to develop a new one, we hope instead that we can contribute to the better understanding of their meaning and the motivation of students' preferences.

4.2. Summary

This paper is the first one in a series in which student preferences in their choice of university are investigated. In this paper we developed a proper data representation partial preference list contained in the students' applications. Based on that data structure different level of aggregation (course, faculty, university) and their analysis and ranking is possible. Several methods of building linear ordered lists from partial lists are explained and compared. As such the obtained student preference lists provide a solid base for further studies to investigate factors influencing students in their choice.

Acknowledgements

This paper was made under the project TÁMOP-4.2.2/B-10/1-2010-0025.

The authors are grateful to András Farkas for the elaborated explanation of linear ordering methods and several hints which proved to be essential in this work. Sincere thanks go to the team of FELVI.hu / Educatio Kht. for their help and authorization of usage of application data. We are indebted to the faculty members of University Pannonia for their helpful comments and suggestion.

References

Alonso S., Chiclana F. Herrera F., Herrera-Viedma E., Alcala-Fdez, J.. Porcel C. (2005). A consistency based procedure to estimate missing pair-wise preference values. tech. rep., Department of Computer Science and Artificial Intelligence, University of Granada, Spain.

Braun H. (1991). On solving travelling salesman problems by genetic algorithms. Lecture Notes in Computer Science, Volume 496, 129-133, DOI: 10.1007/BFb0029743.

Farkas, A., Lancaster, P., Rózsa, P. (2003), Consistency adjustments for pairwise comparison matrices. Numer. Linear Algebra Appl., 10: 689–700. doi: 10.1002/nla.318

Kemeny, J. (1959). Mathematics without numbers. Daedalus 88, 577–591.

Kendall, M. (1962). Rank Correlation Methods. Hafner, New York, NY, USA, 3rd edition

Martí, R., Reinelt, G. (2011). The Linear Ordering Problem: Exact and Heuristic Methods in Combinatorial Optimization. Applied Mathematical Sciences, Volume 175, Springer.

Fedrizzi, M., Giove, S. (2007). Incomplete pairwise comparison and consistency optimization. European Journal of Operational Research 183(1), 303-313

Mitchell, M. (1996). An Introduction to Genetic Algorithms. MIT Press, Cambridge, MA.

Shin, J., Toutkoushian, R., Teichler, U. (2011). Ed.s University Rankings, Theoretical Basis. Methodology and Impacts on Global Higher Education, Springer 2011. Springer.

Tanino, T. (1984). Fuzzy preference orderings in group decision making. Fuzzy Sets and Systems 12, 117-131.

Török, Á. (2009): On the economics of university ranking lists: intuitive remarks on intuitive comparisons. In: Attila,Varga (ed.): Universities, Knowledge Transfer and Regional Development. Geography, Entrepreneurship and Policy. Edward Elgar, Cheltenham, UK – Northampton, MA, USA, 219-242.

van Vught, F.A., Ziegele, F. (Eds.) (2012). Multidimensional Ranking The Design and Development of U-Multirank Series: Higher Education Dynamics, Vol. 37, http://ec.europa.eu/education/higher-education/doc/multirank_en.pdf .