The Use of Collaboration Distance in Scheduling Conference Talks

Jan Pisanski
University of Ljubljana, Faculty of Arts,
E-mail: Jan.Pisanski@ff.uni-lj.si, http://oddelki.ff.uni-lj.si/biblio/oddelek/osebje/pisanski.html

Tomaž Pisanski
University of Primorska, FAMNIT
E-mail: Tomaz.Pisanski@upr.si, https://en.wikipedia.org/wiki/Tomaz Pisanski
ORCID 0000-0002-1257-5376

Keywords: collaboration distance, scheduling, orthogonal partitions

Received: June 15, 2019

Several bibliographic databases offer a free tool that enables one to determine the collaboration distance or co-authorship distance between researchers. This paper addresses a real-life application of the collaboration distance. It concerns somewhat unusual clustering; namely clustering in which the average distances in each cluster need to be maximised. We briefly consider a pair of clusterings in which two cluster partitions are uniform and orthogonal in the sense that in each partition all clusters are of the same size and that no pair of elements belongs to the same cluster in both partitions. We consider different objective functions when calculating the score of the pair of orthogonal partitions. In this paper the Wiener index (a graph invariant, known in chemical graph theory) is used. The main application of our work is an algorithm for scheduling a series of parallel talks at a major conference.

Povzetek: Nekatere bibligrafske zbirke podatkov nudijo orodje, ki za poljubna raziskovalca poišče njuno razdaljo sodelovanja, oz. razdaljo soavtorstva. Članek obravnava konkretno uporabo razdalje sodelovanja. Pri tem gre za nekoliko nenavadno razvrščanje podatkov, pri katerem morajo biti razdalje med elementi skupine čim večje. Na kratko obravnavamo par uniformnih razvrščanj, pri katerem ima vsaka skupina prve komponente z vsako skupino druge komponente natanko en skupen element. Omenimo različne kriterijske funkcije za izračun vrednosti razvrščanj. V praksi uporabimo Wienerjev indeks, ki ga dobro poznamo v kemijski teoriji grafov. Glavna uporaba našega dela je algoritem za razporejanje serije vzporednih predavanj na večji konferenci.

1 Introduction

In this paper we address the use of collaboration distance in solving several practical problems. In particular we apply it to scheduling conference talks in parallel. A problem facing organizers of large conferences where several talks are scheduled in parallel is to avoid simultaneous talks of speakers that may interest the same person, or at least to minimize the number of attendees who have to choose between two interesting talks. Another, somehow complementary task is to schedule similar talks in the same session, preferably in the same lecture room and next to each other. So the main question is, what function one has to take to measure similarity between two speakers. In this paper we will use an objective approach to these ends and simply employ the collaboration distance, information that is readily available in some bibliographic databases.

2 Collaboration graph and collaboration distance

2.1 Collaboration graph

Let V be a list of researchers. This list may be obtained in any manner, but it makes sense to base it on (preferably authority controlled) lists of authors from bibliographic databases. We say that $u, v \in V$ are adjacent: $u \sim v$, if u and v collaborate. Usually, by collaboration we mean that they have written a joint publication in the past. In this sense we consider collaboration to be the same thing as co-autorship. Since $\sim$ is a binary irreflexive, symmetric relation it defines a simple graph $G = (V, \sim)$ that we call the *collaboration graph*[1]. Clearly, one has to specify the data set from which relation $\sim$ can be deduced. Hence G depends on the choice of such a data set.

[1] Here we present the basic model that suffices for our purposes. Note that some studies use reflexive relation signifying that each author collaborates with himself or herself. Also, the graph may be weighted where the weights on the edges represent the number of joint papers between the two authors.

2.2 Collaboration distance

Any connected graph G gives rise to a metric space where the distance $d(u, v)$ between two vertices $u, v \in V$ is defined as the length of the shortest path in G between u and v. If G is disconnected, each of its connected components is a metric space and we let $d(u, v) = \infty$ for vertices in different connected components. For a collaboration graph G the expression $d(u, v)$ is called a *collaboration distance* between authors u and v.

For basics in graph theory, the reader is referred to [6]; for metric spaces, see [7].

2.3 Data sets

It seems the first idea of collaboration graph and collaboration distance appeared as entertainment among mathematicians when measuring how close their research is from the prolific mathematician Pál Erdős. The corresponding collaboration distance is called the Erdős number, and was first formally introduced forty years ago [10]. Scientific investigation of Erdős collaboration graph began in 2000 [5]. Soon it became clear that the same data set can be used for computation of collaboration distance between any two individuals, not only the distance from one particular subject. One can easily define other collaboration graphs, e.g. among movie actors. There is an edge between two actors if and only if they have appeared in the same movie. Collaboration graphs became important in social sciences as prominent examples of social networks. Large social networks exhibit characteristic features of random networks. Modern theory of random networks was born in 1999 [1] when the model was proposed which explains very well the nature of social networks such as collaboration graphs.

Nowadays, two large bibliographic databases covering research in mathematics exist: *MathSciNet* that is run by the American Mathematical Society and *ZbMath*, run by the European Mathematical Society via Springer. Both cover most important publications in mathematics, statistics and theoretical computer science. Each of them contains a tool for calculating the collaboration distance between two authors. In our application the collaboration distances between speakers were taken from ZbMath.

Unfortunately, other important bibliographic databases such as Web of Science, SCOPUS or Google Scholar, do not provide free tools for computing collaboration distance. Slovenia has an excellent research information system SICRIS/COBISS that covers the work of over 15,000 Slovenian scientists. Although it has been analysed with respect to collaboration distance, only summary results in form of scientific papers are available, see e.g. [2, 3, 9, 11]. We strongly believe that a collaboration graph and the corresponding collaboration distance function based on SICRIS should be made available on-line.

3 Selecting optimal orthogonal partitions

Here we present an application of collaboration distance to a sample of individuals.

3.1 Scheduling talks in parallel

Let V be a set of speakers at a scientific conference. Assume each speaker delivers a single talk and that all talks are to be scheduled in parallel in m lecture rooms. Let $n = |V|$. To simplify our task we assume that there are t equal time-slots available and that $n = tm$.[2] Our task is to partition the set of speakers into t groups $U_1, U_2, \ldots, U_t$ such that each group U_i contains m speakers that will speak at the same time. At the same time we want to partition the speakers into m groups $L_!, L_2, \ldots L_{m\cdot}$, assigning each group to a lecture room. In other words we are restricting our search to the pair of *uniform partitions*.

Group	L_1	$\ldots$	$L_{m\cdot}$
U_1	v_{11}	$\ldots$	v_{1m}
U_2	v_{21}	$\ldots$	v_{2m}
$\ldots$	$\ldots$	$\ldots$	$\ldots$
U_t	v_{t1}	$\ldots$	v_{tm}

Table 1: Partitioning the set of speakers into t clusters U_i, representing time slots and an orthogonal partitioning into m clusters L_j , representing lecture rooms.

We would like to choose a partition in which the researchers in each part U work on different topics. A good measure may be collaboration distance.[3] If two researchers have a paper in common they should probably be in different parts. We would like collaboration distances in each group as big as possible. At the same time we would like to have the clusters in the other, orthogonal partition to be as homogeneous as possible. We decided to use a function that is well-known in chemical graph theory, namely, the *Wiener index.*

3.2 The Wiener index of an induced subgraph and clustering

Let G be a connected graph. The Wiener index $W(G)$ is defined as:

$$W(G) = (1/2) \sum_{u \in V} \sum_{v \in V} d(u, v)$$

[2] In more general case when the divisibility condition is violated one could introduce slack or dummy speakers and appropriately define the distances for them.

[3] Any of several other measures, such as citations, keywords, etc, could have been used.

We may restrict this index to a subgraph, induced by $U \subset V$.

$$W(G,U) = (1/2) \sum_{u \in U} \sum_{v \in U} d(u,v)$$

This notion can be found, for instance in [7].

Let $\mathcal{U}$ be a partition of V into t parts of size m, each. The partitioning may be called a *clustering* and each part may be called a *cluster*.

We generalise the notion of the transmission of a vertex in a graph; see [8]. Let v be a vertex, then the sum:

$$w(G,U,v) = \sum_{u \in U} d(v,u)$$

is called the *transmission* of v to U in G. Note that Dobrynin in [8] only considers the case when $U = V$. Given cluster U, the element $u \in U$ with minimal transmission is called a *clustroid* of U. Clustroids are used in several clustering algorithms. However, we will use them only *post festum*.

For a clustering $\mathcal{U}$ define

$$F(\mathcal{U}) = \sum_{U \in \mathcal{U}} W(G,U)$$

We are searching for an admissible partition $\mathcal{U}$ that will maximise $F(\mathcal{U})$. As we show below one may refine this search by adding another, orthogonal criterion.

3.3 Orthogonal clusters and orthogonal partitions

The same data and the same criterion function can be used in the opposite direction, namely to cluster speakers into sections. This means that the talks in the same section will be scheduled consecutively in the same lecture room.

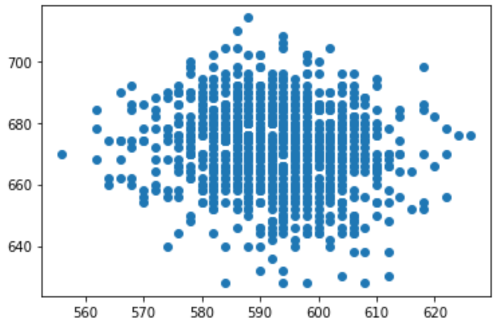

Figure 1: $F(\mathcal{U})$ vs. $F(\mathcal{L})$ for 10000 random permutations π. The optimal results and Pareto frontier can be found in the bottom right.

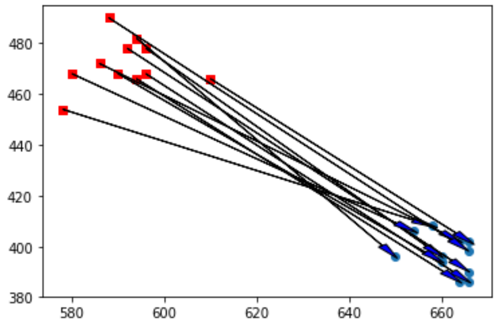

Figure 2: $F(\mathcal{U})$ vs. $F(\mathcal{L})$ for 10 random permutations π each followed by a local optimisation.The initial scores are in top left squares while the optimal scores and Pareto frontier can be found on the bottom right circles. Arrows link each square to the corresponding circle.

In case we want to perform both tasks simultaneously, we may choose to consider *orthogonal partitions*. Two uniform partitions of an mk-set are orthogonal if one has clusters of size k and the other one of size m and no pair of elements belongs to both partitions. In one partition we want to maximize distances while in its orthogonal mate we minimize distances.

Let $c = (\mathcal{U}, \mathcal{L})$ be a pair of orthogonal partitions of V. Let F be defined as above. Define the *score* of $(\mathcal{U}, \mathcal{L})$ to be $F(\mathcal{U}) - F(\mathcal{L})$. Note that each permutation π of V, i.e. $\pi \in Sym(V)$, can be considered as a pair $(\mathcal{U}, \mathcal{L})$. Hence $F(\pi) = F(\mathcal{U}) - F(\mathcal{L})$. We chose the solution to be $\mathrm{argmax}_{\pi \in \mathrm{Sym(V)}} \mathrm{F}(\pi)$.[4]

The task we wanted to solve was the scheduling of 30 invited speakers of the 8th European Congress of Mathematics that is taking place in Portorož, Slovenia in July 2020. The Congress takes place in 5 consecutive days and each day 6 speakers have to deliver their talks in parallel.

In the first attempt we generated 1000 admissible solutions randomly. The results are depicted in Figure 1. We also wrote a program for improving each admissible solution by local optimisation. This improved the quality of the final solution considerably. Figure 2 depicts 10 runs of our algorithm. The top left dots correspond to the randomly generated solutions while the bottom right ones depict the ones, obtained by a sequence of improvements leading to a local minimum. The arrows join each initial solution to the corresponding locally optimal one.

[4] Note that this can be considered also as a multi-criteria optimisation problem with score $(F(\mathcal{U}), -F(\mathcal{L}))$ with Pareto points being candidate solutions.

3.4 Alternative candidates for a score of an orthogonal pair of partitions.

The choice of $F(\pi)$ may not be most suitable for the task.

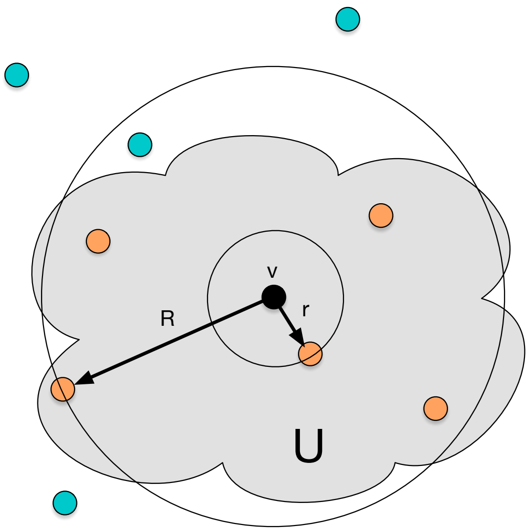

Figure 3: Radius R and the isolation radius r of point v with respect to U. Points in U are in the gray cloud.

We extend the definition of the *radius of a cluster* [12], to the radius of any individual v with respect to the cluster U:

$$R(G,U,v) = \max_{u \in U} d(v,u)$$

Since our clusters are in a sense *anticlusters* as they contain individuals being as far apart as possible, it make sense to define another radius that we call the *isolation radius*

$$r(G,U,v) = \min_{u \in U, d(u,v)>0} d(v,u)$$

measuring the distance to the nearest element in the cluster; see Figure 3.

Note that transmissions measure average distance, while the radius and isolation radius measure maximal and minimal distance, respectively. Also, the *centroid* is a vertex attaining the maximum radius in has been used extensively in data science. We may define *isolation centroid* as the vertex attaining minimal isolation radius.

Since we are already given a distance matrix, data preprocessing is not needed. If needed a method that has all clusters of equal size can be used.

It would be probably interesting to select the pair $(\mathcal{U},\mathcal{L})$ by maximizing the sum of isolation radii in $\mathcal{U}$ and minimizing the sum of radii in $\mathcal{L}$. There are other well-known techniques, such as greedy method or integer programming that should be investigated for this problem.

4 Some further applications of collaboration distance

Collaboration distance can be used as a basis for natural structuring of a given list of researchers using standard clustering methods. We envision several applications of this approach including two that we mention here.

In the first approach one can focus on researchers belonging to a given organization, such as university, institute, faculty, department, project, etc. The internal structure of various universities and institutes could be compared to the collaboration network. Figure 4 is just an illustration of a simple application that gives a very natural stratification of a mathematical department in Slovenia in which three subgroups of researchers are clearly identified. Again, collaboration distances from ZbMath were used. We intend to pursue further studies in this direction.

The second one involves clustering of individuals of a given bibliographic database. Namely, having collaboration graph consisting of all researchers in a given database or country would be very useful. One could use it, in principle, to analyse similarity between various institutions, research groups and scientific disciplines. Various anomalies could be detected and used by policy makers to change the rules in order to avoid it in the future.

While the two mathematics databases (MathSciNet and ZbMath) provide the users with collaborative distance for a given pair of authors, most of the databases in other fields, as well as general databases, do not. This means that additional work must be done by users to find collaboration distances between authors. There are other factors to consider, when calculating Erdős numbers. Firstly, consistent data on the authors is needed, which implies at least consistent spelling of the names, but preferably authority control using consistent identifiers[5]. If this condition is not met, the results will not be appropriate. Next, the range of publications considered for calculation, can have a significant effect on the calculated collaboration distance. For a given pair of researchers their collaboration distance can be, and is, different for different databases. That only a certain subclass of publications is considered, is more or less an arbitrary decision, which is usually a reflection of the scope of a particular database.

One can envision other situations where different distances may be significant. For instance, when selecting referees for a paper one would like to select objective ones, i.e. the ones that are not co-authors of candidates. On the other hand we would like so select individuals who know well the subject, covered in the paper or project under review. This closeness may be measured, for instance by the overlap of keywords used by the two individuals.

Clearly, a *fractional approach* [4] in which the collaboration distance is not measured simply as a distance in the collaboration graph but the number of joint papers shortens the distance accordingly.

[5] One of the most well-know identifiers is ORCID.

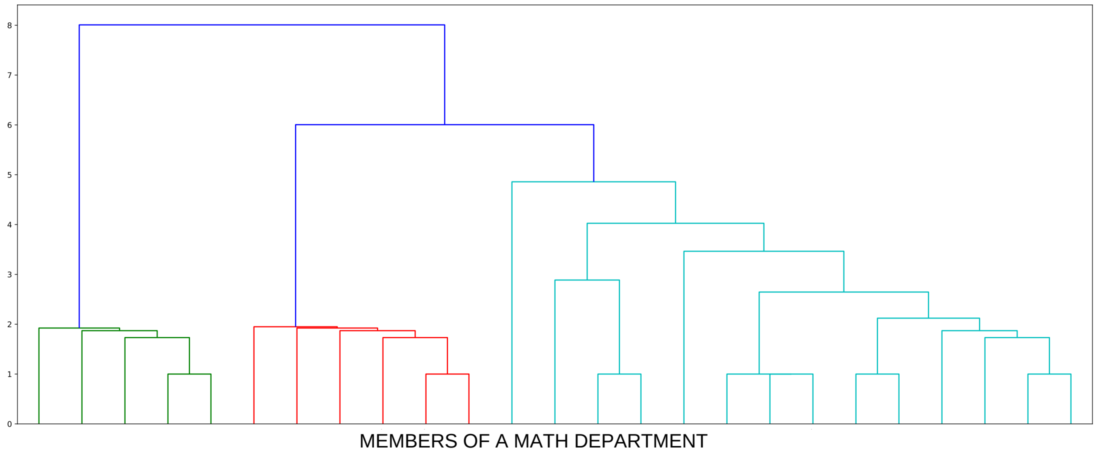

Figure 4: Ward method clustering based on the ZbMath collaboration distance for a department gives a reasonable partition of its members in three groups.

5 Conclusion

The main goal of this paper was to point out that the collaboration distance that is available at some high-quality bilbliographic databases such as MathSciNet and ZbMath is a useful tool that can be applied to a variety of specific problems such as scheduling talks at conferecnes or analysing internal structures of universities, institutes, etc. However, it would be very useful if one could specify the types of edges of the collaboration graphs. For instance, in MathSciNet co-authorship of editorial does not count. It would be useful if the user could choose criteria for inclusion/exclusion of data from the dataset. An important fact may be the time-frame of joint publications. For instance, by looking at recent co-authorships one could easily detect possible conflicts of interest. For other purposes it would be helpful to have information how many co-authors contributed to the edge of the collaboration graph and more generally the number of shortest paths connecting two authors.

Having such a simple tool incorporated into SICRIS would be an important upgrade of the system. One could also look at other measures of similarity, however, it would probably be difficult to get an agreement which ones to include. We would like to stress that we are not doing massive data mining. Our real-life calculations involved rather small data sets. For larger conferences with over 1000 active participants one should perhaps look for methods that would reduce the size of data that is needed to store the distance matrix. It would be interesting to explore how the attendees of a conference choose the talks they attend. In particular, it would be interesting to compare the proposed clustering approach to the manual organization of talks.

Acknowledgement

We would like to thank Vladimir Batagelj and Mark Pisanski for useful advice and fruitful discussion. The work of both referees is gratefully acknowledged. It improved the presentation of results considerably. Work of Jan Pisanski is supported in part by the ARRS grants P5-0361 and J5-8247, while work of Tomaž Pisanski is supported in part by the ARRS grants P1-0294,J1-7051, N1-0032, and J1-9187.

References

[1] A.L Barabási and R Albert. Emergence of scaling in random networks. *Science*, vol. 286 (1999), no. 5439, pp. 509–512.
https://doi.org/10.1126/science.286.5439.509

[2] T. Bartol, K. Stopar, and G. Budimir. Visualization and knowledge discovery in metadata enriched aggregated data repositories harvesting from Scopus and Web of Science. *Information management in the big data era: for a better world : Selected IMCW2015 Papers*. Sun Yat-sen University North: Hacettepe University, 2015. pp 1–5.

[3] T. Bartol, et al. Mapping and classification of agriculture in Web of Science: other subject categories and research fields may benefit. *Scientometrics*, vol. 109 (2016), no. 2, pp. 979–996.
https://doi.org/10.1007/s11192-016-2071-6

[4] V. Batagelj. On Fractional Approach to Analysis of Linked Networks, *arxiv* (2019) https://arxiv.org/abs/1903.00605.

[5] V. Batagelj and A. Mrvar. Some analyses of Erdős collaboration graph. *Social Networks,* vol. 22 (2000),

no. 2, pp. 173–186.
https://doi.org/10.1016/S0378-8733(00)00023-X

[6] J.A. Bondy and U.S.R. Murty. *Graph theory*, (2008) Graduate Texts in Mathematics, 244. Springer, New York.
https://doi.org/10.1007/978-1-84628-970-5

[7] M. M. Deza and E. Deza. *Encyclopedia of distances.* Fourth edition. (2016), Springer, Berlin.
https://doi.org/10.1007/978-3-662-52844-0

[8] A.A. Dobrynin. On 2-connected transmission irregular graphs *Diskretn. Anal. Issled. Oper.*, vol. 25 (2018), no. 4, pp. 5–14.

[9] A. Ferligoj et al. Scientific collaboration dynamics in a national scientific system. *Scientometrics*, vol. 104 (2015), no. 3, pp. 985–1012.
https://doi.org/10.1007/s11192-015-1585-7

[10] C. Goffman. And what is your Erdős number?, *Amer. Math. Monthly*, vol. 76 (1979), p. 791
https://doi.org/10.2307/2317868

[11] L. Kronegger, F. Mali, A. Ferligoj, and P. Doreian. Collaboration structures in Slovenian scientific communities. *Scientometrics*, vol. 90 (2012), no.2, pp. 631–647.
https://doi.org/10.1007/s11192-011-0493-8

[12] J. Leskovec, A. Rajaraman, and J. Ullman. *Mining of Massive Datasets* (2014), Cambridge University Press.
https://doi.org/10.1017/CBO9781139924801

[13] MathSciNet:
https://mathscinet.ams.org/mathscinet/index.html

[14] SICRIS:
https://www.sicris.si/public/jqm/cris.aspx?lang=eng

[15] zbMATH:
https://zbmath.org/