Summary of the PhD thesis

Beyond Dense Subgraphs: Nestedness, Hierarchies, and Community Structures in Complex Networks

IMRE GERA

Supervisors:
András Pluhár, PhD, associate professor
and
András London, PhD, assistant professor

Doctoral School of Computer Science
University of Szeged

Department of Computational Optimization

2024

1 Introduction

In order to analyze and predict the behavior of the various systems that make up our environment, we use models to describe them. In cases where we need to model relationships between entities, we turn to the toolkit of network science, especially when we need to identify groups that share a common property. We often call these groups *communities*. The term "community," however, is not precise, as these groups are often defined as entities with denser internal connections than external ones within a network. Hierarchies also tend to appear together with communities, as seen in a variety of fields, such as social sciences [9], economics [10] and ecology [11]. In the social sciences and economics, we usually see well-defined, formal hierarchies, e.g., with leaders in groups or levels of hierarchies in firms. Sometimes, though, these hierarchies are not explicit, but we still wish to identify them.

In this thesis, I introduce methods that allow detecting overlapping and hierarchical relationships within networks. Nestedness is a famous example that has mostly only been quantified in the literature, not detected as an overlapping community structure [12]. It is a special pattern where entities (e.g., pollinators) with fewer connections are linked to the same group (of plants, for example) as those with more connections. Information about local nested relationships in the network can help us better understand the role of each entity in the network. It also implies a hierarchy, where the subset relationship naturally defines so-called *generalist* and *specialist* entities, that have large or small neighborhoods, respectively. Similarly, knowing what communities (and potentially hierarchies) exist in a network turns out to be a useful piece of information in other domains as well. In portfolio optimization, clusters of assets that are highly correlated are usually best avoided, as investing in them increases the risk of the portfolio. Here, methods have been introduced that use hierarchical clustering to perform portfolio selection [13, 14], however, the Markowitz model [15], which relies on quadratic programming, can also be improved.

The thesis presents new community detection methods and adaptations

of hierarchical clustering algorithms for special use cases. Contributions of this work include heuristic and exact overlapping community detection algorithms for finding fully nested subgraphs in networks, and specializations of hierarchical clustering algorithms to detect nestedness and to find closely related assets in the stock market to aid diversification in the investment process.

The dissertation consists of two major parts. Following a brief introduction to the concepts discussed in **Chapter 1**, **Part I** presents overlapping community detection algorithms for nestedness. **Chapter 2** proposes an edge-based heuristic for finding overlapping nested subgraphs in a network. Following in these footsteps, **Chapter 3** introduces an exact algorithm for discovering all maximal nested subgraphs in a network. In **Part II**, I focus on the problem of finding (disjoint) clusters using hierarchical clustering to solve different problems where a known cluster structure is beneficial. In **Chapter 4**, I adapt two hierarchical clustering approaches for detecting nestedness. **Chapter 5** presents the use of hierarchical clustering to improve the risk estimation of the Markowitz portfolio selection model and compares it against new, hierarchical clustering-based portfolio selection algorithms.

2 Nestedness as an overlapping community structure

Nestedness is a property of networks that imposes stiff structural constraints. In a fully nested graph, the vertices (in the case of bipartite graphs, the vertices of each class) can be ordered such that the neighborhood of a lower degree vertex is a subset of that of a higher degree vertex.

The property also implies a hierarchy, due to the relationship of the neighborhoods. In the chains, vertices with fewer neighbors are often referred to as *specialists*, while vertices with large neighborhoods are called *generalists*. Our goal was to detect the overlapping fully nested subgraphs, and potentially also identify generalist and specialist vertices within the communities. We also wanted to develop methods that do not require graphs to be bipartite, since the definition of nestedness can be applied to

non-bipartite graphs as well [16]. In order to measure nestedness at the local level, we use the following metric:

$$\mathrm{nest}(i,j) = \frac{|N(i) \cap N(j)|}{\min\{|N(i)|, |N(j)|\}}, \tag{1}$$

where $N(i)$ is the neighborhood of vertex i.

First, in **Chapter 2** of the thesis, a heuristic algorithm is presented that uses the edges of a graph to detect overlapping fully nested subgraphs. The algorithm has a parameter that allows fine-tuning its performance-accuracy trade-off. We have evaluated the algorithm on random and real-world bipartite networks and showed that the size of the communities detected by the algorithm correlates with the *discrepancy* nestedness measure. We also measured the runtime and the average community sizes as a function of the threshold parameter. We found that a large fraction of nested communities are detected in the first few iterations of the algorithm, making the parameter an effective tool for increasing performance while maintaining relatively high accuracy. This is shown in Figure 1.

Chapter 3 presents an overlapping community detection algorithm for nestedness that relies on the construction of an auxiliary nestedness graph that allows us to extract the communities and also infer the role of each node in those communities. Additionally, we introduced a method that can create bipartite graphs with known ground truth nested communities, allowing us to test the community detection algorithm. Using the output of the detection algorithm, we defined a new metric (*vertex presence*) to measure graph-level nestedness. We also tested the algorithm on real-world bipartite and non-bipartite networks. We found that the ecological bipartite networks showed some degree of nestedness, with some networks being highly or even fully nested. This was not true in the case of non-bipartite networks, where there were large amounts of small fully nested subgraphs, leading to the conclusion that the networks were not nested. One such example is the Florentine families network seen in Figure 2.

I have published open-source reference implementations for both algorithms [17].

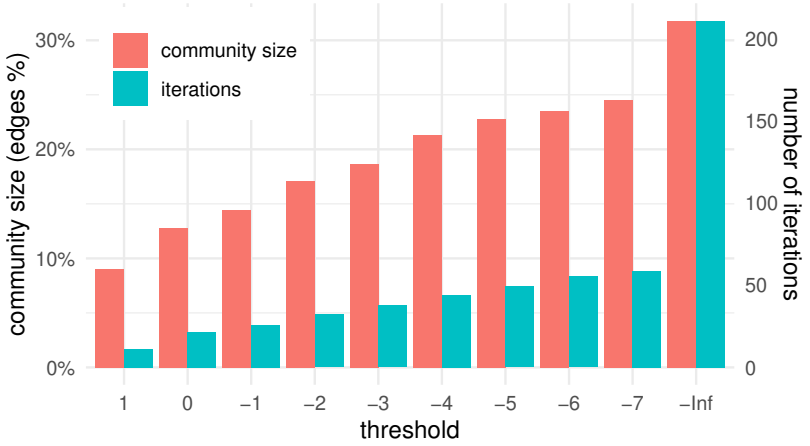


Figure 1: *Average community size and number of iterations performed for each n_1 threshold value on the host-parasite networks.*

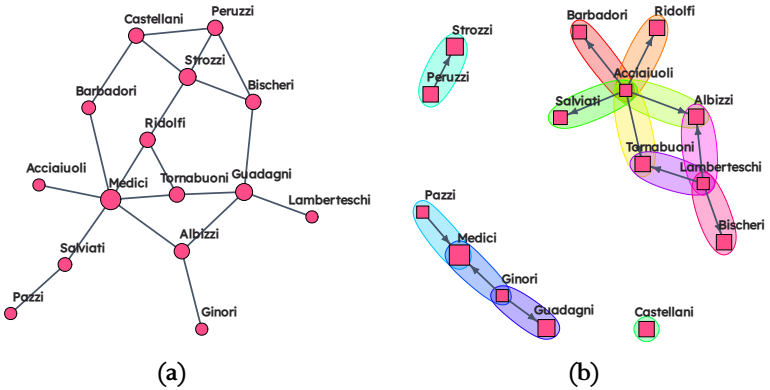


Figure 2: *The Florentine families network and its nestedness community graph. In this network, nestedness is preferably avoided, as having only a subset of another family's neighbors may be highly disadvantageous in an information spreading situation. Only some central families, such as the Medici, Albizzi, or Strozzi, are generalists (with only incoming edges). The nestedness of the network is low.*

3 Applications of hierarchical clustering

While overlapping community detection can reveal more information about the interaction of the different groups in a network through overlaps, clustering is sometimes still the preferred approach. In cases where we want to focus on placing entities in their dominant group, or when comparing groupings on the same graph across algorithms, disjoint clusters are much easier to handle. Hierarchical clustering creates an approach that is halfway between traditional clustering and overlapping community detection methods. It constructs a merge tree representing the hierarchy of the nodes and clusters, where every horizontal slice (level) is a clustering of the network. The tree has n clustering levels, where n is the number of vertices in the network. Between two adjacent levels, two clusters are merged into one or one cluster is split into two, depending on the approach. This makes hierarchical clustering more versatile than traditional clustering algorithms, since its output encodes additional information that enables us to make further adjustments to the results.

In various problems, where networks can be used as a representation, identifying groups of similarly behaving entities is key to solving the problem itself, either directly or indirectly. In **Part II**, I explore one problem in each category: in **Chapter 4**, I continue to work with nestedness, where I use hierarchical clustering to find disjoint fully nested subgraphs in networks, and in **Chapter 5**, I use hierarchical clustering to improve the performance of a portfolio selection model. In the case of nestedness, it is our direct goal to find fully nested groups of vertices, and among them, we want to find a clustering where every cluster is a fully nested subgraph, but the number of clusters is as small as possible. However, in portfolio selection, we want to *avoid* investing heavily in stocks (the vertices) that behave similarly.

Chapter 4 introduces adaptations of two hierarchical clustering approaches for finding disjoint fully nested subgraphs of a network. We adapt a bottom-up (agglomerative) and a top-down (divisive) algorithm, the latter being a modification of the Girvan-Newman algorithm [18], a well-known method for performing divisive hierarchical clustering. To select the appro-

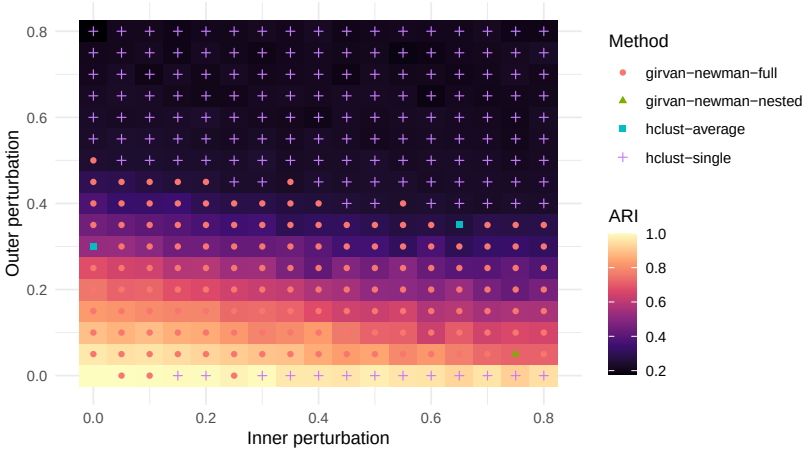


Figure 3: *Best method based on ARI values for each p_1 and p_2 value on the synthetic networks.*

priate clustering level, we introduced new metrics derived from the results of the algorithms that allow measuring the nestedness of both a nested clustering and the whole graph. To evaluate the stability of the methods, we created synthetic bipartite networks with known nestedness structures and rewired their edges with given probabilities. Of the algorithms, the top-down one performed better at lower permutation probabilities, and the bottom-up approach was better at higher ones, as seen in Figure 3. Furthermore, we also tested the algorithms on real-world bipartite and non-bipartite networks. The experiments confirmed that the traditional benchmark non-bipartite networks we analyzed in **Chapter 3** show little nestedness overall. We also showed that the bottom-up approaches found the largest fully nested clusters. A comparison of how nested the clusters are on average at each step on the real-world bipartite network dataset is shown in Figure 4. The algorithms are also available as open source code [19].

Chapter 5 introduces hierarchical clustering algorithms for filtering covariance matrices in the Markowitz portfolio selection model [15]. The

model relies on covariance matrices to assess risk, but these matrices are sensitive to estimation errors. However, to get a clearer picture of how the market behaved in the recent past, we need to use a smaller sample size. The "noise" introduced by the estimation can be reduced by filtering methods. Here, we use hierarchical clustering to implement such filtering methods by creating a graph from the covariance matrix, building a minimal spanning tree-like structure with hierarchical clustering, and using the clustering results to replace the covariances.

We also compared our methods with new portfolio selection algorithms that are based purely on hierarchical clustering and do not rely on covariance matrices [13, 14]. For our experiments, we used two real-world datasets from the assets of the Budapest Stock Exchange (BSE) and the assets of the Standard and Poor's 500 (S&P 500) index. A summary of the returns and risks on the S&P 500 dataset is shown in Figure 5.

Our results show that the filtering algorithms were effective in reducing both the realized risk and the risk estimation errors, while the new methods realized higher returns in exchange for higher risks.

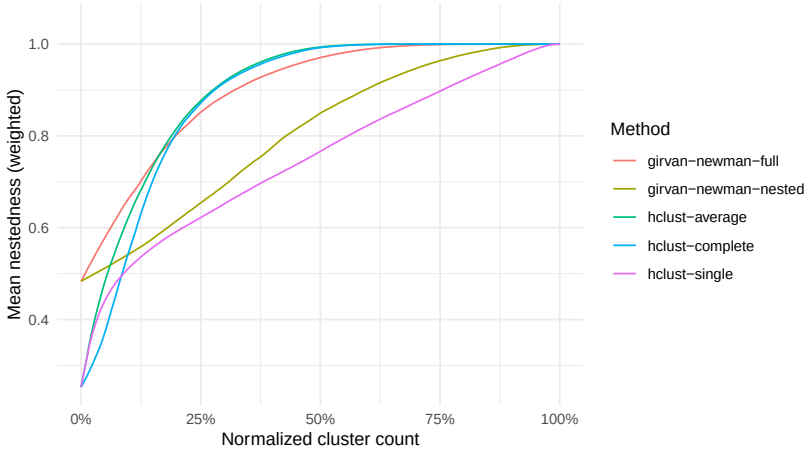
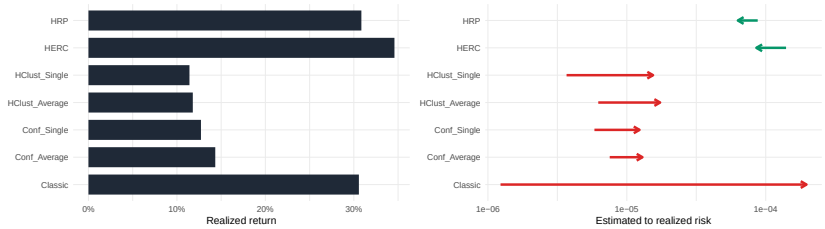


Figure 4: *Average of mean nestedness values across all graphs of the Web of Life dataset in the function of normalized cluster counts, for all clustering algorithms.*



(a) (b)

Figure 5: *Realized returns and changes between estimated and realized risks on the S&P 500 dataset. The arrows point towards the realized risk and their lengths correspond to the magnitude of over- or underestimation – if red, the method underestimated the risk, if green, it overestimated it.*

4 Contributions of the thesis

In the **first thesis group**, the contributions are related to detecting overlapping nested community structures. Detailed discussion can be found in Chapters 2 and 3.

I/1. I proposed an edge-based, adjustable heuristic algorithm for detecting overlapping nested communities.

I/2. I introduced an exact algorithm to derive a nestedness graph from networks. With this graph, all overlapping communities can be detected.

I/3. I defined metrics from the output of the overlapping nested community detection algorithms to quantify graph-level nestedness.

I/4. I showed that the new algorithms (unlike most previously found in the literature) do not require graphs to be bipartite.

I/5. I showed that common non-bipartite benchmark graphs show little, while some bipartite ecological graphs show large amounts of nestedness.

In the **second thesis group**, the contributions are related to the applications of hierarchical clustering. Detailed discussion can be found in Chapters 4 and 5.

II/1. I adapted two hierarchical clustering approaches to detect disjoint fully nested subgraphs in networks.

II/2. I created a unified approach to compare top-down and bottom-up algorithms.

II/3. I defined metrics based on the output of the nested hierarchical clustering algorithms to quantify graph-level nestedness.

Table 1: *Correspondence between the thesis points and my publications.*

Publication	Thesis point									
	I/1	I/2	I/3	I/4	I/5	II/1	II/2	II/3	II/4	II/5
[P1]									●	●
[P2]									●	●
[P3]									●	●
[P4]		●	●	●	●					
[P5]						●	●	●		
[P6]	●		●	●	●					
[P7]									●	●
[P8]									●	●

II/4. I introduced hierarchical clustering-based filtering methods to improve the performance of the Markowitz portfolio selection model by reducing noise.

II/5. I compared multiple hierarchical clustering-based solutions (including filtering algorithms and models) on the portfolio selection problem using real-world datasets, showing that the filtering algorithms are effective in reducing risk estimation errors, while the new methods are capable of achieving higher returns.

Table 1 summarizes the relation between the thesis points and the corresponding publications.

The author's publications on the subjects of the thesis

Journal publications

[P1] London, András, Gera, Imre, and Bánhelyi, Balázs. "Markowitz Portfolio Selection Using Various Estimators of Expected Returns and Filtering Techniques for Correlation Matrices". In: *Acta Polytechnica Hungarica* 15.1 (Jan. 2018), pp. 217–229. DOI: 10.12700/aph.15.1.2018.1.13

[P2] Gera, Imre és London, András. "Gráf alapú dimenzióredukciós heurisztikák részvénypiaci korrelációs mátrixokra". *Alkalmazott Matematikai Lapok* 37.2 (2020. júl.), 211–224. old. DOI: 10.37070/AML.2020.37.2.06

[P3] Gera, Imre és London, András. "Hierarchikus klaszterezés és a portfólió-kiválasztás probléma". *SZIGMA Matematikai-közgazdasági folyóirat* 53.1 (2022), 73–88. old.

[P4] Gera, Imre and London, András. "Detecting and generating overlapping nested communities". In: *Applied Network Science* 8.1 (Aug. 2023). ISSN: 2364-8228. DOI: 10.1007/s41109-023-00575-2

[P5] Gera, Imre and London, András. "Recovering Nested Structures in Networks: An Evaluation of Hierarchical Clustering Techniques". In: *Journal of Complex Networks* (2024). DOI: 10.1093/comnet/cnae039, *(accepted for publication)*

Full papers in conference proceedings

[P6] Gera, Imre, London, András, and Pluhár, András. "Greedy algorithm for edge-based nested community detection". In: *2022 IEEE 2nd Conference on Information Technology and Data Science (CITDS)*. 2022, pp. 86–91. DOI: 10.1109/CITDS54976.2022.9914051

Further related publications

[P7] Gera, Imre, Bánhelyi, Balázs, and London, András. "Testing the markowitz portfolio optimization method with filtered correlation matrices". In: *Proceedings of the 19th International Multiconference INFORMATION SOCIETY - IS 2016*. Institut Jožef Stefan, 2016, pp. 44–47

[P8] Gera, Imre and London, András. "Portfolio selection based on a configuration model and hierarchical clustering for asset graphs". In: *Proceedings of the MATCOS*. vol. 19. 2019

Bibliography

[9] Redhead, Daniel and Power, Eleanor A. "Social hierarchies and social networks in humans". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 377.1845 (Jan. 2022). ISSN: 1471-2970. DOI: 10.1098/rstb.2020.0440.

[10] Radner, Roy. "Hierarchy: The Economics of Managing". In: *Journal of Economic Literature* 30.3 (1992), pp. 1382–1415. ISSN: 00220515.

[11] Huxham, Mark, Raffaelli, Dave, and Pike, Alan. "Parasites and Food Web Patterns". In: *The Journal of Animal Ecology* 64.2 (Mar. 1995), pp. 168–176. ISSN: 0021-8790. DOI: 10.2307/5752.

[12] Mariani, Manuel Sebastian et al. "Nestedness in complex networks: observation, emergence, and implications". In: *Physics Reports* 813 (2019), pp. 1–90. DOI: 10.1016/j.physrep.2019.04.001.

[13] Raffinot, Thomas. "Hierarchical clustering-based asset allocation". In: *The Journal of Portfolio Management* 44.2 (2017), pp. 89–99. DOI: 10.3905/jpm.2018.44.2.089.

[14] Raffinot, Thomas. "The Hierarchical Equal Risk Contribution Portfolio". In: *SSRN Electronic Journal* (Aug. 2018). ISSN: 1556-5068. DOI: 10.2139/ssrn.3237540.

[15] Markowitz, Harry. "Portfolio selection". In: *The Journal of Finance* 7.1 (1952), pp. 77–91.

[16] London, András, Martin, Ryan R, and Pluhár, András. "Graph clustering via generalized colorings". In: *Theoretical Computer Science* 918 (2022), pp. 94–104. DOI: 10.1016/j.tcs.2022.03.023.

[17] Gera, Imre. *Overlapping community detection algorithm for nestedness*. 2023. URL: https://github.com/Hanziness/r-nested-comms/tree/v0.2.

[18] Newman, M. E. J. and Girvan, M. "Finding and evaluating community structure in networks". In: *Phys. Rev. E* 69.2 (2004), p. 026113. DOI: 10.1103/PhysRevE.69.026113.

[19] Gera, Imre. *Hierarchical clustering algorithms for nested community detection*. 2024. URL: https://github.com/Hanziness/r-nested-comms/tree/v0.3.

5 Összefoglalás

Az értekezés új, speciális közösségkereső eljárásokat ismertet. A módszerekben közös, hogy olyan közösségszerkezeteket keresnek, amelyekben hierarchikus viszonyok is fellelhetők. A bemutatott algoritmusok két fő témakörre bonthatók: átfedő közösségkereső és hierarchikus klaszterező eljárások, így a munka is két fő részből áll.

Az **első rész** olyan átfedő közösségeket kereső algoritmusokat mutat be, amelyek teljesen egymásba ágyazott részgráfokat azonosítanak hálózatokban. A 2. fejezetben először egy paraméterezhető heurisztikus algoritmust ismertetek, amely él-alapon keres átfedő, teljesen egymásba ágyazott közösségeket. A 3. fejezetben egy egzakt algoritmust mutatok be, amely minden maximális egymásba ágyazott közösséget felfed a hálózatokban. Az algoritmussal együtt definiálok egy új metrikát (*vertex presence*) is, amellyel gráf szinten mérhető az egymásba ágyazottság. Megállapítottuk, hogy a vizsgált ökológiai hálózatokban magasabb, míg a hagyományos közösségkeresési referenciagráfokban alacsony az egymásba ágyazottság. A szakirodalomban tudtunkkal fellelhető algoritmusokkal ellentétben a mi eljárásaink nem csak páros gráfokon működnek.

A **második rész** hierarchikus klaszterező algoritmusok alkalmazásait mutatja be. A 4. fejezetben két megközelítés több változata is megtalálható, amelyekkel gráfok egymásba ágyazottsága deríthető fel (nem átfedő) klaszterek formájában. Az algoritmusok közül az alulról-felfelé (*bottom-up*) módszerek találták meg a legnagyobb teljesen egymásba ágyazott klasztereket.

Az 5. fejezet hierarchikus klaszterezésen alapuló kovarianciamátrix-szűrési eljárásokat mutat be a Markowitz portfóliómodell teljesítményének javítására. A szűrési eljárásokat új, hierarchikus klaszterezésen alapuló portfóliókiválasztási eljárásokkal is összevetettük. Megállapítottuk, hogy a szűrési eljárások hatékonyabbak voltak a kockázatbecslés hibájának csökkentésében, míg az új eljárások magasabb hozamokat kínáltak, magasabb kockázatok mellett.

Nyilatkozat

Gera Imre "Beyond Dense Subgraphs: Nestedness, Hierarchies, and Community Structures in Complex Networks" című PhD disszertációjában a következő eredményekben Gera Imre hozzájárulása volt a meghatározó:

- A **2. fejezetben** felhasznált [P6] publikációban megjelent kutatás esetén: algoritmus megtervezése és megvalósítása, kísérletek eredményeinek értelmezése és vizualizációja.
- A **3. fejezetben** felhasznált [P4] publikációban megjelent kutatás esetén: algoritmus megtervezése és megvalósítása, tesztgráf-generáló algoritmus megtervezése és megvalósítása, *vertex presence* metrika definiálása, kísérletek kiértékelése és eredmények vizualizációja.
- A **4. fejezetben** felhasznált [P5] publikációban megjelent kutatás esetén: algoritmusok adaptálásának megtervezése és implementációja, szintetikus gráfgeneráló algoritmus megtervezése, egymásba ágyazottsági metrikák definiálása, eredmények kiértékelése és vizualizációja.
- Az **5. fejezetben** felhasznált [P1, P2, P3] publikációkban megjelent kutatás esetén: módszerek implementációja, adathalmazok összeállítása, kísérletek megtervezése és kiértékelése.

Ezek az eredmények Gera Imre PhD disszertációján kívül más tudományos fokozat megszerzésére nem használhatók fel.

Szeged, 2024. szeptember 10.

Gera Imre	Pluhár András	London András
jelölt	témavezető	témavezető

Az Informatika Doktori Iskola vezetője kijelenti, hogy jelen nyilatkozatot minden társszerzőhöz eljuttatta, és azzal szemben egyetlen társszerző sem emelt kifogást.

Szeged, 2024. 09. 11.

SZEGEDI TUDOMÁNYEGYETEM INFORMATIKA DOKTORI ISKOLA SZEGED

Dr. Jelasity Márk
Informatika Doktori Iskola vezetője