Hierarchical organisation in complex systems

Final Report

1 Introduction

Hierarchical organisation is a widespread feature of a large variety of systems studied in natural- and social sciences. Former empirical studies on hierarchy are ranging from the transcriptional regulatory network of Escherichia coli [1], the dominant-subordinate hierarchy among crayfish [2], the leader-follower network of pigeon flocks [3, 4] and the rhesus macaque kingdoms [5], through neural networks [6] and technological networks [7], to social interactions [8, 9, 10], urban planning [11, 12], ecological systems [13, 14], and evolution [15, 16]. Although hierarchy is often easily perceived, it can be difficult to quantify according to the numerous alternative hierarchy detection methods and hierarchy measures available in the scientific literature [1, 17, 18, 19, 20, 21, 22].

A hierarchy is usually depicted as a directed acyclic graph (DAG), where the nodes are layered in different levels, with links pointing from nodes in higher levels towards nodes in lower levels. However, we can distinguish between hierarchies of different types. In a flow hierarchy, the links can represent the chain of command, the flow of information, etc., and in general, they indicate that nodes found in lower levels are influenced by their in-neighbours positioned in higher levels. On the other hand, in a nested hierarchy (also called inclusion hierarchy or containment hierarchy) the links represent an "is a sub-category of" type of relation, as nested hierarchies are obtained by recursively aggregating items into larger and larger groups, resulting in a structure where higher-level groups consist of smaller and more specific components [23]. Prominent examples of nested hierarchy are given by library classification systems and biological classification.

The aim of the project was to study the hierarchical organisation of complex systems from new perspectives, and to continue the former research of project members related to tagging systems [24, 25], ontologies [26] and community finding in complex networks [27, 28]. The obtained results were mainly published in peer reviewed scientific journals (we shall refer to these as [J1], [J2], etc. in the following) and in on case as a peer reviewed conference abstract (we shall refer to that as [C1]). The publications supported by the grant are provided at the end of the report in lists separated from the rest of the references. In the following we give a short summary of our findings divided into sections according to the main focus of the obtained results.

2 Achievements in tag hierarchy extraction [J1, J2, J3]

A very interesting problem related to nested hierarchies and classification is given by the automated extraction of nested hierarchies from folksonomies and collaborative tagging systems [29, 30, 31, 32]. The association of tags to various on-line contents have became widespread, as various tags may indicate the topic of news-portal feeds and blog post, the genre of films or music records on file sharing portals, or the kind of goods offered in Web stores. These tags usually serve as keywords, providing a rough description of the given entity, helping the users in a fast decision whether the given article, film, etc. is of interest or not. Since the tags appearing in these on-line platforms are usually free words chosen by the author or owner of the given object, they are almost never organised into a pre-defined hierarchy of categories and sub-categories [26, 33, 34]. Furthermore, in many tagging systems like Flickr, CiteUlike or Delicious the tagging process is collaborative, as in principle an unlimited number of users can tag photos, Web pages, etc., with free words [35, 36, 37]. In order to highlight this collaborative nature, the arising set of free tags and associated objects are often referred to as folksonomies.

One of our first main achievements during the project was to provide a complete framework for automated tag hierarchy extraction based on tag occurrence statistics in folksonomies and tagging systems [J1]. This included the introduction of two novel tag hierarchy extraction algorithms, which were thoroughly tested on several real systems, including a large data set of annotated proteins, tagged by their known functions. Here the "exact" hierarchy between the protein functions can be also looked up in the Genome Ontology. The similarity between this pre-defined hierarchy and the reconstructed hierarchies obtained from our algorithms was very encouraging. Furthermore, the hierarchies obtained for large samples of tagged photos from Flickr and tagged films from the IMDb database were seemingly meaningful, following the common sense by and large. (No pre-defined exact hierarchies were given for these systems).

Another important part of the our framework was the development of a computer generated benchmark system providing a tunable and controllable test bed for tag hierarchy extraction methods [J1]. The basic idea of this approach is the simulation of the association of tags to virtual objects based on a given input hierarchy between the tags. According to our results, by changing the parameters of the simulated tagging process we are capable of generating both "easy" and "difficult" data sets for the tag hierarchy reconstructing algorithms. In addition, we have also introduced several quality measures for evaluating the similarity between a reconstructed hierarchy, (obtained from a tag hierarchy extraction method), and an exact hierarchy.

The developed tag hierarchy extraction framework was later applied to keywords associated to scientific publications in the Web of Science [J2]. In this study we treated the keywords given by the authors and the keywords given by the repository separately, resulting in two alternative hierarchies. Interestingly, the keywords were also pre-organised by the Web of Science into 3 types, from the very general to the very special. When comparing the two extracted hierarchies to the 3-level classification given by the Web of Science we found a very good correspondence, as the specificity of the keywords according to the Web of Science categorisation was increasing with the level depth in the extracted hierarchies [J2]. In addition, the hierarchy between the author given tags and the hierarchy of the repository given tags were also compared to each other. According to the results, the top of the hierarchies showed a relatively high similarity, whereas the bottom of the hierarchies (where the specificity of the tags was high) turned out to be quite dissimilar. This is in accordance with the fact that for each individual paper in the Web of Science the tags given by the repository are aimed to be complementary to the keywords given by the authors.

We also carried out a study focusing on the hierarchy between tags associated to news feeds in on-line news portals such as the Spiegel Online, The Guardian, The New York Times and The Australian [J3]. The extracted hierarchies suggested that the tags appearing in the different news portals were organised to different degrees. Our analysis revealed that Guardian has an extra intermediate level of organisation at certain locations. A further very interesting result is that the number of connected components in the DAGs conveys information about the extent of organisation in the data: the Spiegel and Guardian have $\mathcal{O}(1)$ components and are quite organised, the New York Times has a few dozen components and breaks the world into independent pieces, and the Australian has $\mathcal{O}(100)$ components which are barely informative at all. A similar picture was emerging from the comparison between the frequencies of tags in Google News and their centrality score in the tag-tag co-appearance graphs [J3]. The correlation was quite strong in case of the Spiegel and the Guardian, medium for the New York Times, and almost equivalent to the totally random case for the Australian.

3 Results related to communities [J4, J5, J6]

Over the last 15 years, the complex network approach in general has become ubiquitous in describing and modelling complex systems [38, 39]. One of the most intensively studied area in this field is devoted to communities (also called as modules, clusters, cohesive groups, etc.) [27, 40, 41, 42, 43]. The concept of nested communities, where smaller, strongly connected modules are embedded in larger and larger shells of more loose communities, is naturally very closely related to containment hierarchy.

Clique percolation is a popular community finding approach, in which nested communities can be obtained in a natural way, just by changing the k-clique size used for detecting the communities [27]. The usual "rule of the thumb" for selecting the optimal value for k (and if the network is weighted, selecting an optimal value for a link weight threshold w^* used for throwing away unimportant links) is to tune the system close to the critical point of k-clique percolation. Relating to that, we examined the behaviour of the quality of community partitions at the critical point using several different modularity measures [J4]. According to our results on several real networks, the studied modularities show a maximum close to the critical point, justifying the mentioned criteria for the optimal parameter settings in the Clique Percolation Method.

We have also made some progress related to the problem of tagging of communities in a study focusing on the changing role of past scientific publications [J5]. Although a scientific paper cannot be changed after publication, its "meaning" to the reader and the categorisation of its content may change over time. In order to examine how scientific ideas (represented by groups of papers) evolve, we prepared the yearly co-citation network between publications appearing in the Web of Science, identified time evolving communities using the Clique Percolation Method, and assigned topic tags to all groups. The main results of the study were that the tag "Wos multidisciplinary" was over-represented among cutting edge ideas, and that we can find a number of scientific papers moving from one community (corresponding to a certain topic) to another one (devoted to some other scientific field) [J5]. We found also that several groups of scientific papers change their focus with time in the sense, that the most typical keywords of the groups are changing.

Tagging of tightly connected sub-graphs in a network has several possible applications. We have used this concept to help physicians in better understanding complex and rare diseases [J6]. We have analysed the clinical pathways of a population with relapsing polychondritis, which is a rare autoimmune disease with many manifestations. Based on the co-occurring diagnoses we have identified several distinct symptoms. We have also shown, that by analysing the time evolution of the patient paths that are tagged by ICD-10 diagnosis codes and ATC codes of consumed medicaments one can distinguish between severity levels [J6]. These levels of severity were confirmed by independent data, where the length of hospitalisation was used as indicator of severity.

4 Hierarchical properties of scientific citation networks [J7, C1, J8, J9]

Beside the (undirected) co-citation network mentioned in the previous Section, the directed citation network between scientific publications provides another fascinating complex system. Due to the appearance of new papers, this network is subject to constant evolution. In one of our studies we examined the hierarchical structure of time dependent sub-networks consisting of publications belonging to a given scientific field [J7]. According to the results, the majority of these networks follows a universal trend towards a highly hierarchical state, and the observed variations over the fields in most cases can be accounted for a difference in either their maturity (distance in time from the "birth" of the field), or a difference in the characteristic time according to which they approach their stationary state. Furthermore, the alteration in the behaviour of the sub-networks turned out to be closely related also to the degree of specialisation of the fields [J7].

Citations play also a very important role in scientometrics when comparing the quality, novelty and impact of scientific publications. Probably the most widely used article level metric is the accumulated number of citations a paper collects over time. However, this can have a number of drawbacks, e.g., when comparing papers that appeared decades ago (and had a lot of time to collect citation) to newly appearing articles. We have proposed a new measure for quantifying the novelty of a manuscript based on the frequency of citations between the papers citing the given document and the cited other articles in the document [C1]. The main idea behind this metric is that if the majority of the papers citing the given article in the mean time have references also to the cited papers in the document, then the given article does not play a significant role in producing new knowledge. In contrast, if the later coming new papers cite only the given document, but not the previous publications cited in the document, then the manuscript is likely to contain an important new contribution to the field. We have analysed six different data sources (two databases of scientific publications, three online Wikipedias and one patent database). According to our results the above novelty measure does not correlate with the plain citation count of the papers [C1].

Another drawback of using the citation count as an article level metric is that the average frequency of citations shows large variations across different scientific fields (making a comparison between papers from far away fields unfair). However, since the majority of papers listed in current large scientific publication repositories are tagged by keywords or field categories, we may naturally restrict the comparison of citation counts to papers belonging to the same category or having the same tags. In a short study we gave a brief overview of a few alternative article level metrics aimed at providing a more fair comparison between papers, and also examined a few basic statistics of topic groups obtained by simply collecting publications having a given keyword in the repository of the Web of Science [J8].

By continuing the extension of the width of scientific topics, we may move from the previous groups of papers marked by the same keywords to much broader groups, corresponding to papers appearing in the same scientific journal. Quantifying the impact, prestige and novelty of a journal is yet again, a non-trivial problem [44, 45, 46, 47, 48], and according to comparison studies between journal impact measures, scientific impact seems to be a multi-dimensional construct that cannot be adequately measured by any single indicator [49, 50, 51]. Motivated by that, we carried out a study focusing on the possibilities for organising scientific journals into hierarchies, which can be considered as a first step in the direction of multi-dimensional characterisation of journal impact [J9]. We succeeded in constructing both a flow hierarchy, (relevant from the point of view of information spreading), and a nested hierarchy, (organising the journals according to the scientific fields), using the publication data from the Web of Science. According to the results, our nested hierarchy provides a natural tool for the visualisation of the intricate nested and overlapping relations between scientific fields, making the navigation within different scientific fields and sub-fields very simple, being equivalent to navigating in the different branches of the hierarchy. Furthermore, based on the position of a journal in the hierarchy, we gain immediate information on its standing within its particular field. According to that, we can select those journals with which we can make a fair comparison, and we can exclude journals in far away branches from any comparing study. In addition, similarly to judging the position of a journal within its specific field (a local branch), we can also judge the standing of this sub-field in a larger scientific domain, (a main branch), and so on, and thereby, compare the ranking of the different scientific fields and sub-fields (each being composed of multiple journals) [J9].

5 Development of a novel hierarchy measure [J10]

One of the key problems related to hierarchies is the quantification of the importance of hierarchical organisation in the structure of the network representing the interactions or connections between the fundamental units of the studied system. Although a number of notable hierarchy measures were already introduced in the scientific literature [18, 21, 22, 52, 53], their vast majority is treating all directed acyclic graphs as already maximally hierarchical. This means that e.g., a directed chain, a directed tree and a directed star (where all the nodes have a single incoming link from a central hub) are considered equally hierarchical by these methods. In contrast, the general intuition of a hierarchy is usually corresponding to a multi-level pyramidal structure, with levels becoming wider and wider as we descend from the root towards the bottom. On the one hand this way the top nodes in the hierarchy can reach most of the network in a very effective way, i.e., via paths of average length scaling as $\ln N$, where N denotes the number of nodes. On the other hand, in this structure all nodes can have a treatable number of direct subordinates (opposed to e.g., the star configuration).

Motivated by that, we developed a novel hierarchy measure based on random walks on the network that gives higher scores to directed trees compared to directed chains or directed stars [J10]. A notable further feature of our approach is that in the thermodynamic limit the hierarchy measure of regular trees is converging to a well defined limit, depending only on the branching number. Our method turned out to be computationally very effective, and tests on several real world networks provided very intuitive results, e.g., the trophic levels obtained from our approach on a food web were highly consistent with former results from ecology [J10].

Publications in peer reviewed scientific journals

[J1] Tibély, G., Pollner, P., Vicsek, T. & Palla, G. Extracting Tag Hierarchies. *PLoS ONE* **8**, e84133 (2013).

[J2] Tibély, G., Pollner, P. & Palla, G. Comparing the hierarchy of author given tags and repository given tags in a large document archive. *The European Physical Journal Special Topics* **225**, 2025 (2016).

[J3] Tibély, G., Sousa-Rodrigues, D., Pollner, P. & Palla, G. Comparing the Hierarchy of Keywords in On-Line News Portals. *PLoS ONE* **11**, e0165728 (2016).

[J4] Tóth, B., Vicsek, T. & Palla, G. Overlapping modularity at the critical point of k-clique percolation. *J. Stat. Phys.* **151**, 689–706 (2013).

[J5] Orosz, K., Farkas, I. J. & Pollner, P. Quantifying the changing role of past publications. *Scientometrics* **108**, 829–853 (2016).

[J6] Horváth, A., Páll, N., Molnár, K., Kováts, T., Surján, G., Vicsek, T. & Pollner, P. A nationwide study of the epidemiology of relapsing polychondritis. *Clinical Epidemiology* **8**, 211–230 (2016).

[J7] Mones, E., Pollner, P. & Vicsek, T. Universal hierarchical behavior of citation networks. *J. Stat. Mech.* **2014**, P05023 (2014).

[J8] Szántó-Várnagy, Á., Pollner, P., Vicsek, T. & Farkas, I. J. Scientometrics: Untangling the topics. *National Science Review* **1**, 343–345 (2014).

[J9] Palla, G., Tibély, G., Mones, E., Pollner, P. & Vicsek, T. Hierarchical networks of scientific journals. *Palgrave Communications* **1**, 15016 (2015).

[J10] Czégel, D. & Palla, G. Random walk hierarchy measure: What is more hierarchical, a chain, a tree or a star? *Scientific Reports* **5**, 17994 (2015).

Publication as a peer reviewed conference abstract

[C1] Szántó-Várangy, Á., Pollner, P. & Farkas, I. J. Measuring originality in knowledge networks. *Lect. Notes Comput. Sc.* **9197**, 156799 (2015).

References

[1] Ma, H. W., Buer, J. & Zeng, A. P. Hierarchical sructure and modules in the Escherichia coli transcriptional regulatory network revealed by a new top-down approach. *BMC Bioinformatics* **5**, 199 (2004).

[2] Goessmann, C., Hemelrijk, C. & Huber, R. The formation and maintenance of crayfish hierarchies: behavioral and self-structuring properties. *Behav. Ecol. Sociobiol.* **48**, 418—428 (2000).

[3] Nagy, M., Akos, Z., Biro, D. & Vicsek, T. Hierarchical group dynamics in pigeon flocks. *Nature* **464**, 890—893 (2010).

[4] Nagy, M. *et al.* Context-dependent hierarchies in pigeons. *Proc. Natl. Acad. Sci. USA* **110**, 13049—13054 (2013).

[5] Fushing, H., McAssey, M. P., Beisner, B. & McCowan, B. Ranking network of captive rhesus macaque society: A sophisticated corporative kingdom. *PLoS ONE* **6**, e17817 (2011).

[6] Kaiser, M., Hilgetag, C. C. & Kötter, R. Hierarchy and dynamics of neural networks. *Front. Neuroinform.* **4**, 112 (2010).

[7] Pumain, D. (ed.) *Hierarchy in Natural and Social Sciences*, vol. 3 of *Methodos Series* (Springer Netherlands, Dodrecht, The Netherlands, 2006).

[8] Guimerà, R., Danon, L., Díaz-Guilera, A., Giralt, F. & Arenas, A. Self-similar community structure in a network of human interactions. *Phys. Rev. E* **68**, 065103 (2003).

[9] Pollner, P., Palla, G. & Vicsek, T. Preferential attachment of communities: The same principle, but a higher level. *Europhys. Lett.* **73**, 478–484 (2006).

[10] Valverde, S. & Solé, R. V. Self-organization versus hierarchy in open-source social networks. *Phys. Rev. E* **76**, 046118 (2007).

[11] Krugman, P. R. Confronting the mystery of urban hierarchy. *J. Jpn. Int. Econ.* **10**, 399—418 (1996).

[12] Batty, M. & Longley, P. *Fractal Cities: A Geometry of Form and Function* (Academic, San Diego, 1994).

[13] Hirata, H. & Ulanowicz, R. Information theoretical analysis of the aggregation and hierarchical structure of ecological networks. *J. Theor. Biol.* **116**, 321—-341 (1985).

[14] Wickens, J. & Ulanowicz, R. On quantifying hierarchical connections in ecology. *J. Soc. Biol. Struct.* **11**, 369—378 (1988).

[15] Eldredge, N. *Unfinished Synthesis: Biological Hierarchies and Modern Evolutionary Thought* (Oxford Univ. Press, New York, 1985).

[16] McShea, D. W. The hierarchical structure of organisms. *Paleobiology* **27**, 405––423 (2001).

[17] Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. & Barabási, A.-L. Hierarchical organization of modularity in metabolic networks. *Science* 1551–1555 (2002).

[18] Trusina, A., Maslov, S., Minnhagen, P. & Sneppen, K. Hierarchy measures in complex networks. *Phys. Rev. Lett.* **92**, 178702 (2004).

[19] Sales-Pardo, M., Guimerá, R., Moreira, A. A. & Amaral, L. A. N. Extracting the hierarchical organization of complex systems. *Proc. Natl. Acad. Sci. USA* **104**, 15224–15229 (2007).

[20] Clauset, A., Moore, C. & Newman, M. E. J. Hierarchical structure and the prediction of missing links in networks. *Nature* **453**, 98–101 (2008).

[21] Corominas-Murtra, B., Goñi, J., Solé, R. V. & Rodríguez-Caso, C. On the origins of hierarchy in complex networks. *Proc. Natl. Acad. Sci. USA* **110**, 13316—13321 (2013).

[22] Mones, E., Vicsek, L. & Vicsek, T. Hierarchy measure for complex networks. *PLoS ONE* **7**, e33799 (2012).

[23] Wimberley, E. T. *Nested ecology. The place of humans in the ecological hierarchy* (John Hopkins University Press, Baltimore, MD, 2009).

[24] Palla, G., Farkas, I. J., Pollner, P., Derényi, I. & k, T. V. Fundamental statistical features and self-similar properties of tagged networks. *New Journal of Physics* **10**, 123026 (2008).

[25] Pollner, P., Palla, G. & Vicsek, T. Clustering of tag-induced subgraphs in complex networks. *Physica A* **389**, 5887–5894 (2010).

[26] Tibély, G., Pollner, P., Vicsek, T. & Palla, G. Ontologies and tag-statistics (2011). ArXiv:1201.1085.

[27] Palla, G., Derényi, I., Farkas, I. & Vicsek, T. Uncovering the overlapping community structure of complex networks in nature and society. *Nature* **435**, 814–818 (2005).

[28] Palla, G., Barabási, A.-L. & Vicsek, T. Quantifying social group evolution. *Nature* **446**, 664–667 (2007).

[29] Heymann, P. & Garcia-Molina, H. Collaborative creation of communal hierarchical taxonomies in social tagging systems. Technical Report, Stanford InfoLab (2006). URL `http://ilpubs.stanford.edu:8090/775/`.

[30] Plangprasopchok, A., Lerman, K. & Getoor, L. A probabilistic approach for learning folksonomies from structured data. In *Fourth ACM International Conference on Web Search and Data Mining (WSDM)*, 555–564 (2011).

[31] Schmitz, P. Inducing ontology from flickr tags. In *Proc. of Collaborative Web Tagging Workshop at the 15th Int. Conf. on World Wide Web (WWW)* (2006).

[32] Damme, C. V., Hepp, M. & Siorpaes, K. Folksontology: An integrated approach for turning folksonomies into ontologies. *Social Networks* **2**, 57–70 (2007).

[33] Mika, P. Ontologies are us: A unified model of social networks and semantics. In *In International Semantic Web Conference, Vol. 3729*, 522–536 (2005).

[34] Spyns, P., Moor, A. D., Vandenbussche, J. & Meersman, R. From Folksologies to Ontologies: How the Twain Meet. In *In Proceedings of OTM Conferences (1)*, 738–755 (2006).

[35] Cattuto, C., Loreto, V. & Pietronero, L. Semiotic dynamics and collaborative tagging. *Proc. Natl. Acad. Sci. USA* **104**, 1461–1464 (2007).

[36] Lambiotte, R. & Ausloos, M. Collaborative tagging as a tripartite network. *Lect. Notes in Computer Sci.* **3993**, 1114–1117 (2006).

[37] Cattuto, C., Barrat, A., Baldassarri, A., Schehr, G. & Loreto, V. Collective dynamics of social annotation. *Proc. Natl. Acad. Sci. USA* **106**, 10511–10515 (2009).

[38] Albert, R. & Barabási, A.-L. Statistical mechanics of complex networks. *Rev. Mod. Phys.* **74**, 47–97 (2002).

[39] Mendes, J. F. F. & Dorogovtsev, S. N. *Evolution of Networks: From Biological Nets to the Internet and WWW* (Oxford Univ. Press, Oxford, 2003).

[40] Girvan, M. & Newman, M. E. J. Community structure in social and biological networks. *Proc. Natl. Acad. Sci. USA* **99**, 7821–7826 (2002).

[41] Rosvall, M. & Bergstrom, C. T. An information-theoretic framework for resolving community structure in complex networks. *Proc. Natl. Acad. Sci. USA* **104**, 7327 (2007).

[42] Fortunato, S. Community detection in graphs. *Physics Reports* **486**, 75–174 (2010).

[43] Fortunato, S. & Hric, D. Community detection in networks: A user guide. *Physics Reports* **659**, 1–44 (2016).

[44] Bordons, M., Fernandez, M. T. & Gomez, I. Advantages and limitations in the use of impact factor measures for the assessment of research performance. *Scientometrics* **53**, 195—206 (2002).

[45] Braun, T., Gläanzel, W. & Schubert, A. A hirsch-type index for journals. *Scientometrics* **69**, 169–173 (2006).

[46] Egghe, L. Theory and practice of the g-index. *Scientometrics* **69**, 131–152 (2006).

[47] Bollen, J., de Sompel, H. V., Smith, J. & Luce, R. Toward alternative metrics of journal impact: a comparison of download and citation data. *Inform. Process. Manag.* **41**, 1419–1440 (2005).

[48] Leydesdorff, L. Betweenness centrality as an indicator of the interdisciplinarity of scientific journals. *J. Am. Soc. Inform. Sci.* **58**, 1303—1319 (2007).

[49] Bollen, J., de Sompel, H. V., Hagberg, A. & Chute, R. A principal component analysis of 39 scientific impact measures. *PLoS ONE* **4**, e6022 (2009).

[50] Glänzel, W. The application of characteristic scores and scales to the evaluation and ranking of scientific journals. *J. Inf. Sci.* **37**, 40–48 (2011).

[51] Kaur, J., Radicchi, F. & Menczer, F. Universality of scholarly impact metrics. *J. Informetr.* **7**, 924—-932 (2013).

[52] Krackhardt, D. Graph theoretical dimensions of informal organizations. In Carley, K. & Prietula, M. (eds.) *Computational organization theory*, 89–111 (Lawrence Erlbaum Associates, Hillsdale, NJ, 1994).

[53] Luo, J. & Magee, C. L. Detecting evolving patterns of self-organizing networks by flow hierarchy measurement. *Complexity* **16**, 53–61 (2011).