On Graph Entropy Measures for Knowledge Discovery from Publication Network Data

Andreas Holzinger[1], Bernhard Ofner[1], Christof Stocker[1],
André Calero Valdez[2], Anne Kathrin Schaar[2], Martina Ziefle[2],
and Matthias Dehmer[3]

[1] Medical University Graz, A-8036 Graz, Austria
Institute for Medical Informatics, Statistics & Documentation,
Research Unit Human-Computer Interaction
{a.holzinger,b.ofner,c.stocker}@hci4all.at
[2] Human-Computer Interaction Center, RWTH Aachen University, Germany
{calero-valdez,schaar,ziefle}@comm.rwth-aachen.de
[3] Institute for Bioinformatics and Translational Research, UMIT Tyrol, Austria
matthias.dehmer@umit.at

Abstract. Many research problems are extremely complex, making interdisciplinary knowledge a necessity; consequently cooperative work in mixed teams is a common and increasing research procedure. In this paper, we evaluated information-theoretic network measures on publication networks. For the experiments described in this paper we used the network of excellence from the RWTH Aachen University, described in [1]. Those measures can be understood as graph complexity measures, which evaluate the structural complexity based on the corresponding concept. We see that it is challenging to generalize such results towards different measures as every measure captures structural information differently and, hence, leads to a different entropy value. This calls for exploring the structural interpretation of a graph measure [2] which has been a challenging problem.

Keywords: Network Measures, Graph Entropy, structural information, graph complexity measures, structural complexity.

1 Introduction and Motivation for Research

Tradition in the history of science emphasizes the role of the individual genius in scientific discovery [3]. However, there is an ongoing trend away from such an individual based model of scientific advance towards a networked team model [4]. Teams can bring-in greater collective knowledge; however, the most convincing factor is, that multidisciplinary teams are able to maintain an integration and appraisal of different fields, which often provides an atmosphere to foster different perspectives and opinions; and this often stimulates novel ideas and enables a fresh look on methodologies to put these ideas into business [5].

This is mainly due to the fact that many research problems, e.g. in the life sciences, are highly complex, so that know-how from different disciplines is necessary. Cooperative work in cross-disciplinary teams is thus of increasing interest.

A. Cuzzocrea et al. (Eds.): CD-ARES 2013, LNCS 8127, pp. 354–362, 2013.

Consequently, mixed-node publication network graphs can be used to get insights into social structures of such research groups, but elucidating the elements of cooperation in a network graph reveals more than simple co-authorship graphs, especially as a performance metric of interdisciplinarity [1].

However, before we can select measures to improve communication effectiveness or interpersonal relationships, it is necessary to determine which factors contribute to the interdisciplinary success and furthermore what constitutes interdisciplinary success. Moreover, it is quite important to understand the measures in depth, i.e., what kind of structural information they detect [6], [7], [8], [9].

2 Methods and Materials

As in previous work [1] we use a mixed node graph in order to analyze publication behavior. We create a reduced mixed node public network to demonstrate the research efforts of an interdisciplinary research cluster at the RWTH Aachen University. Typically bibliometric data is visualized using co-authorship graphs, leaving out the element of the interaction (i.e. the publication). The use of mixed node publication network graphs allows a graph to contain more information (than a co-authorship graph) and can easily be reduced to one by using an injective mapping function. This type of graph allows fast human analysis of interdisciplinarity by explicating the authors tension between his discipline and his (possibly interdisciplinary) publications. When visualized properly this graph will match the users mental model, which is important in recognition tasks [10]. In our particular case we use the reduced Graph G_r.

2.1 Construction of the Network Graph

The network graph G_r is constructed equally as in previous work [1] with two node types. A node in this case represents either an author (A-Node) or a publication (P-Node). Nonetheless both node types (i.e. vertices) are not regarded as differently from a graph theory point of view. We define the two sets representing authors and publications as follows:

$$A = \{a \mid \text{a is author in cluster of excellence at RWTH}\} \tag{1}$$

$$P = \{p \mid p \text{ is a publication funded by the cluster written by any } a \in A\} \tag{2}$$

We also define two vertex-mappings f_a and $f_p$and two sets of vertices V_1 and V_2 as follows:

$$f_a : A \rightarrow V_1, f_a(a) = v; a \in A \wedge v \in V_1 \tag{3}$$

$$f_p : P \rightarrow V_2, f_p(p) = v; p \in P \wedge v \in V_2 \tag{4}$$

$$\text{with } V_1 \cap V_2 = \varnothing \tag{5}$$

These function represent mappings of authors and publications to vertices, and if inverted finding the "meaning" of a vertice. We define the sets E as all edges between authors and publications, when an author has written a publication:

$$E = \{e \mid e = (v_1, v_2), v_1 \in V_1 \wedge v_2 \in V_2 \wedge f_a^{-1}(v_1) \text{ is author of } f_p^{-1}(v_2)\} \quad (6)$$

We use the graph definition for Gr for our analyses:

$$G_r = \{(V, E) \mid V = V_1 \cup V_2\} \quad (7)$$

This bipartite graph can be visualized using standard graph visualization tools. In order to enable analysis by a human person the graph needs be lain out visually. Graph visualization is done with both Gephi [11] and D3JS (www.d3js.org). In this case 2D-spatial mapping is performed by using force-based algorithms.

For our visualization we set the size of nodes according to their corresponding degree and applied a grayscale color scheme based on betweenness centrality (see Fig. 1). Fig. 2 shows the two different node sets. White nodes denote authors, gray nodes denote publications. Both graphs share the same layout (Force Atlas 2, linlog mode, no overlap, scaling=0.3, gravity=1) and node sizes (min=10, max=50). Nodes sizes were chosen as node degrees. Authors with more publications are bigger, as well as publications with more authors are bigger.

Using force-based algorithms has the following consequences for graph visualization:

- All nodes are attracted to the center, but repel each other.
- All nodes that are connected by an edge attract each other (i.e. an author and his publication).

This allows the following visual conclusions:

- Two A-Nodes are spatially closer if they publish together (triangle inequality).

2.2 On Graph Entropies

The open source graph visualization tool Gephi allows for several different graph analyses of network graphs. Traditionally these are used with social network graphs (i.e. co-authorship graphs). Interpretation of graph statistics must be reevaluated for mixed node graphs. Graph statistics that are of interest in regard to publication networks are:

- Network Entropies have been developed to determine the structural information content of a graph [7], [2]. We have to mention that the term network entropy cannot be uniquely defined. A reason for this is that by using Shannon's entropy [12], [13], [14] the probability distribution cannot be assigned to a graph uniquely. In the scientific literature, two major classes have been reported [7], [6], [15]:

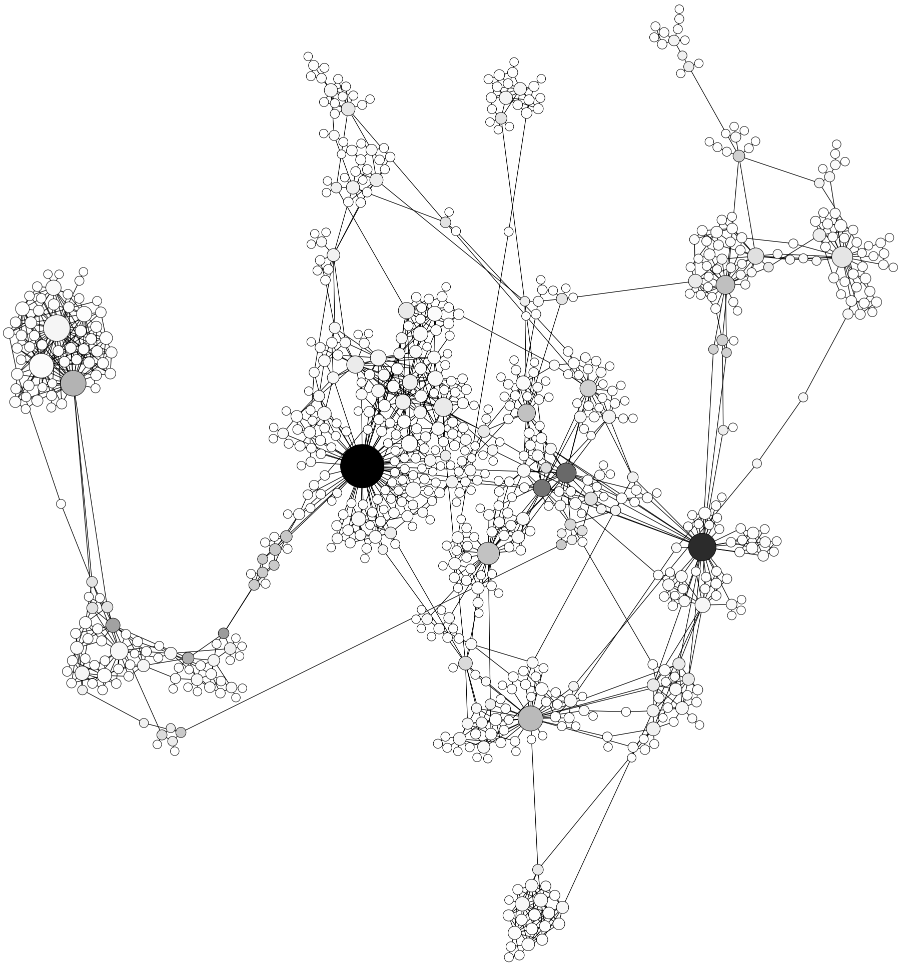

Fig. 1. G_r of publication data of the excellence network from the RWTH Aachen. The node size shows the node degree whereas the node color shows the betweenness centrality. Darker color means higher centrality.

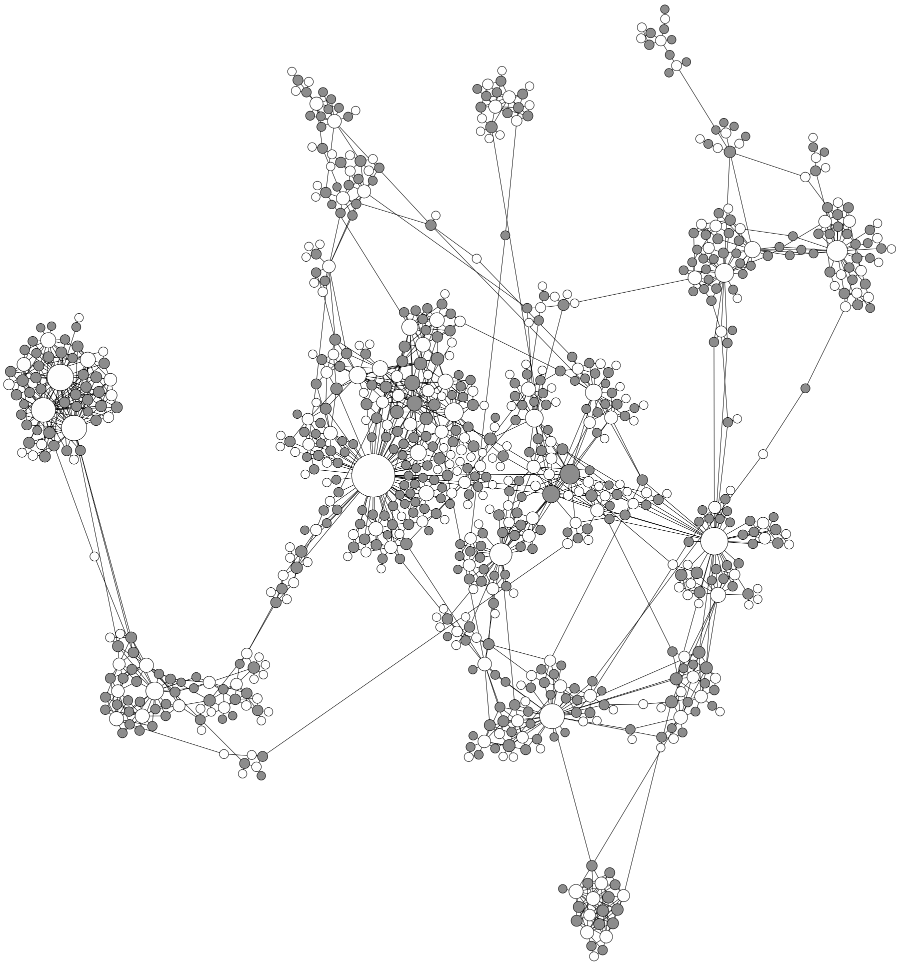

Fig. 2. G_r of publication data of the excellence network from the RWTH Aachen. The node size shows the betweenness centrality. White nodes denote authors, gray nodes denote publications.

1. Information-theoretic measures for graphs which are based on a graph invariant X (e.g., vertex degrees, distances etc.) and an equivalence criterion [13]. By starting from an arbitrary graph invariant X of a given graph and an equivalence criterion, we derive a partitioning. Thus, one can further derive a probability distribution. An example thereof is to partition the vertex degrees (abbreviated as $\delta(v)$) of a graph into equivalence classes, i.e., those classes only contain vertices with degree $i = 1, 2, ..., \max \delta(v)$, see e.g. [8].
2. Instead of determining partitions of elements based on a given invariant, Dehmer [6] developed an approach which is based on using so called information functionals. An information functional f is a mapping which maps sets of vertices to the positive reals. The main difference to partition-based measures (see previous item) is that we assign probability values to every individual vertex of a graph (and not to a partition), i.e.,

$$p^f(v_i) := \frac{f(v_i)}{\sum_{j=1}^{|V|} f(v_j)} \tag{8}$$

As the probability values depend on the functional f, we infer a family of graph entropy measures

$$I_f(G) := -\sum_{i=1}^{|V|} p^f(v_i) \log p^f(v_i) \tag{9}$$

$|V|$ is the size of the vertex set of G. Those measures have been extensively discussed in [8].

Evidently, those information-theoretic graph measures can be interpreted as graph complexity measures.

The Graph G_r contains 796 nodes split into 323 authors and 473 publications linked by 1677 edges. Applying the Gephi graph analysis reveals the following statistics. The graph shows an average degree of 4.214 and a network diameter of 23. The average path length is 7.805 and graph density is .005. The graph only contains a single connected component.

3 Evaluation of the Graph by Using Network Entropies

In this section, we evaluate some information-theoretic network measures (graph entropies) on the given Excellence Network. To start, we briefly characterize this network by stating some graph-theoretical measures.

We evaluated the following graph entropies:

- A partition-based graph entropy measure called *topological information content* based on vertex orbits due to [13].

- Parametric graph entropies based on a special information functional f due to Dehmer [6]. The information functional we used is

$$f(v_i) := \sum_{k=1}^{\rho(G)} c_k |S_k(v_i, G)|, \text{ with } c_k > 0 \tag{10}$$

summing the product of both the size of the k-sphere (i.e. the amount of nodes in G with a distance of k from v_i given as $|S_k(v_i, G)|$) and arbitrary positive correction coefficients c_k for all possible k from 1 to the diameter of the graph G. The resulting graph entropies have been defined by

$$I_f := -\sum_{i=1}^{|V|} p^f(v_i) \log p^f(v_i) \tag{11}$$

- Network entropy due to [16].
- Graph entropy based on the ER model for modeling random graphs [17].

Table 1. Calculated graph entropies

Method	**Symbol**	**Graph Entropy**
Topological information content [13]	I_{mowsh}	9.031485
Parametric graph entropies [6]	I_{dehm}	9.6258
Network entropy due to [16]	I_{valv}	0.3095548
Graph entropy based on the ER model [17]	I_{wang}	15090.71

We can note that the used graph entropies evaluate the complexity of our network differently. Here we will explore this problem with in illustrative exmaple, namely by considering the measures $I_{\text{mowsh}} < I_{\text{dehm}}$. In this context, the inequality $I_{\text{mowsh}} < I_{\text{dehm}}$ can be understood by the fact those entropies have been defined on different concepts.

As mentioned, I_{mowsh} is based upon the automorphism group of a graph and, therefore, can be interpreted as a symmetry measure. This measure vanishes if all vertices are located in only one orbit. By contrast, the measure is maximal ($= \log_2(|V|)$) if the input graph equals the so-called identity graph; that means all vertex orbits are singleton sets. In our case, we obtain $I_{\text{mowsh}} = 9.0315 < \log_2(796) = 9.6366$ and conclude that according to the definition of I_{mowsh}, the excellence network is rather unsymmetrical.

Instead, the entropy I_{dehm} characterizes the diversity of the vertices in terms of their neighborhood, see [7]. The higher the value of I_{dehm}, the less topologically different vertices are in the graph and, finally, the higher is the inner symmetry of our excellence network. Again, maximum entropy for our network equals $\log_2(796) = 9.6366$. Based on the fact that for the complete graph K, $I_{\text{dehm}}(K_n) = \log(n)$ holds, we conclude from the result $I_{\text{dehm}} = 9.6258$ that the excellence network is highly symmetrical and connected and could theoretically be obtained by deleting edges from K_{796}.

The interpretation of the results for I_{valv} and I_{wang} can be done similarly by arguing based on their definitions.

4 Discussion

Different entropy measure deliver different results because they are based on different graph properties. When using the aforementioned entropy measures in a mixed-node publication graph measures of symmetry I_{dehm} (based on vertex neighborhood diversity) or I_{mowsh} (based on the graph automorphism) deliver different measures of entropy. Interpreted we could say, that authors/publications are similar in regard to their neighborhoods (i.e. authors show similar publication behavior, publications show similar author structures) but the whole graph shows low measures of automorphism-based symmetry to itself. This could mean authors or publications can not be exchanged for one another without changing basic properties of the graph. But since authors and publications are used in the same vertex set there are also implications of interpretation between these sets. For example a graph isomorphisms that maps vertices from V_1 to V_2 should not be included in the measure, because they are not intelligible from an interpretation point of view. New measures of entropy specialized for mixed-node graphs are required to accurately measure graph properties in such graphs.

5 Conclusion

In this paper, we evaluated information-theoretic network measures on publication networks. In our case, we used the excellence network from the RWTH Aachen, described in [1]. Those measures can be understood as graph complexity measures which evaluate the structural complexity based on the corresponding concept.

A possible useful interpretation of these measures could be applied in understanding differences in subgraphs of a cluster. For example one could apply community detection algorithms and compare entropy measures of such detected communities. Relating these data to social measures (e.g. balanced score card data) of sub-communities could be used as indicators of collaboration success or lack thereof, as proposed in [18] and [19].

Nonetheless we see that it is challenging to generalize such results towards different measures as every measure captures structural information differently and, hence, leads to a different entropy value. This calls for exploring the *structural interpretation* of a graph measure [2] which has been a challenging problem.

Acknowledgments. We cordially thank the Aachen House of Production for allowing us to analyze their publication data. This research was partially funded by the Excellence Initiative of the German federal and state governments.

References

1. Calero Valdez, A., Schaar, A.K., Ziefle, M., Holzinger, A., Jeschke, S., Brecher, C.: Using mixed node publication network graphs for analyzing success in interdisciplinary teams. In: Huang, R., Ghorbani, A.A., Pasi, G., Yamaguchi, T., Yen, N.Y., Jin, B. (eds.) AMT 2012. LNCS, vol. 7669, pp. 606–617. Springer, Heidelberg (2012)

2. Dehmer, M.: Information theory of networks. Symmetry 3(4), 767–779 (2011)
3. Merton, R.: The Matthew Effect in Science: The reward and communication systems of science are considered. Science 159(3810), 56–63 (1968)
4. Wuchty, S., Jones, B., Uzzi, B.: The increasing dominance of teams in production of knowledge. Science 316(5827), 1036–1039 (2007)
5. Holzinger, A.: Successful Management of Research Development. BoD–Books on Demand (2011)
6. Dehmer, M.: Information processing in complex networks: Graph entropy and information functionals. Appl. Math. Comput. 201(1-2), 82–94 (2008)
7. Dehmer, M., Varmuza, K., Borgert, S., Emmert-Streib, F.: On entropy-based molecular descriptors: Statistical analysis of real and synthetic chemical structures. Journal of Chemical Information and Modeling 49, 1655–1663 (2009)
8. Dehmer, M., Mowshowitz, A.: A history of graph entropy measures. Inf. Sci. 181(1), 57–78 (2011)
9. Holzinger, A., Stocker, C., Bruschi, M., Auinger, A., Silva, H., Gamboa, H., Fred, A.: On Applying Approximate Entropy to ECG Signals for Knowledge Discovery on the Example of Big Sensor Data. In: Huang, R., Ghorbani, A.A., Pasi, G., Yamaguchi, T., Yen, N.Y., Jin, B. (eds.) AMT 2012. LNCS, vol. 7669, pp. 646–657. Springer, Heidelberg (2012)
10. Calero Valdez, A., Ziefle, M., Alagöz, F., Holzinger, A.: Mental models of menu structures in diabetes assistants. In: Miesenberger, K., Klaus, J., Zagler, W., Karshmer, A. (eds.) ICCHP 2010, Part II. LNCS, vol. 6180, pp. 584–591. Springer, Heidelberg (2010)
11. Salotti, J., Plantevit, M., Robardet, C., Boulicaut, J.F.: Supporting the Discovery of Relevant Topological Patterns in Attributed Graphs (December 2012), Demo Session of the IEEE International Conference on Data Mining (IEEE ICDM 2012)
12. Shannon, C.E.: A mathematical theory of communication. Bell System Technical Journal 27 (1948)
13. Mowshowitz, A.: Entropy and the complexity of graphs: I. An index of the relative complexity of a graph 30, 175–204 (1968)
14. Holzinger, A., Stocker, C., Peischl, B., Simonic, K.M.: On using entropy for enhancing handwriting preprocessing. Entropy 14(11), 2324–2350 (2012)
15. Mowshowitz, A., Dehmer, M.: Entropy and the complexity of graphs revisited. Entropy 14(3), 559–570 (2012)
16. Solé, R., Valverde, S.: Information Theory of Complex Networks: On Evolution and Architectural Constraints. In: Ben-Naim, E., Frauenfelder, H., Toroczkai, Z. (eds.) Complex Networks. Lecture Notes in Physics, vol. 650, pp. 189–207. Springer, Heidelberg (2004)
17. Ji, L., Bing-Hong, W., Wen-Xu, W., Tao, Z.: Network entropy based on topology configuration and its computation to random networks. Chinese Physics Letters 25(11), 4177 (2008)
18. Jooss, C., Welter, F., Leisten, I., Richert, A., Schaar, A., Calero Valdez, A., Nick, E., Prahl, U., Jansen, U., Schulz, W., et al.: Scientific cooperation engineering in the cluster of excellence integrative production technology for high-wage countries at rwth aachen university. In: ICERI 2012 Proceedings, pp. 3842–3846 (2012)
19. Schaar, A.K., Calero Valdez, A., Ziefle, M.: Publication network visualisation as an approach for interdisciplinary innovation management. In: IEEE Professional Communication Conference (IPCC) (2013)