To appear in the first issue of *Bolletino di Archaeologia On-line* 2009

The Network of Subject Co-Popularity in Classical Archaeology

Maximilian Schich[1] with César Hidalgo[2], Sune Lehmann[1] and Juyong Park[1]

[1] Center for Complex Network Research, Northeastern University, Boston/MA (http://www.barabasilab.com)
[2] Center for International Development, Harvard University, Cambridge/MA (http://www.cid.harvard.edu)

Abstract

In this paper we map a number of complex network properties of *Archäologische Bibliographie.* In particular we explain the general structure of thematic subdivisions, as well as the (co-)popularity of specific subjects in publications within the field of Classical Archaeology as recorded by the bibliographic database since 1956. In order to map these phenomena we use methods and tools from the science of complex networks. Our results provide a proof of concept for further investigations, which will heighten our understanding of archaeological datasets and lead the way to a new big picture of the discipline.*

Introduction

As a spin-off from the former *Subject Catalogue of the German Archaeologic Institute in Rome* run by *Stiftung Archäologie*, *Archäologische Bibliographie* catalogues new acquisitions of archaeological literature by the American, British, French, German, and Spanish Institutes in Rome.[1] At the time of analysis in March 2008 it contained 426.108 titles (monographs, articles, and other publications) of which 373.191 are connected to 45.924 classification criteria via 617.518 classification links. Currently, the database grows by 25.000 titles a year, which is nearly eight times its growth rate in 1956 and two and a half times its rate in 2001, when it was run by the *German Archaeologic Institute*.

Method

In our analysis of *Archäologische Bibliographie* we use methods from the science of complex networks – a multidisciplinary effort, investigating the relationship patterns that emerge in social, biological, economic and technological systems.[2] We do this by interpreting *Archäologische Bibliographie* as a *network* whose *nodes* are individual database records and whose *links* are database references.[3]

In this paper we deal with two particular types of nodes – *classification criteria* and *publications* – and three types of links: (i) the parent link, which connects classification criteria among each other forming the so called *tree of subject headings*, i.e. the controlled vocabulary of Archäologische Bibliographie; (ii) the *classification link*, which connects publications to their respective classification criteria, forming a *bipartite network*, i.e. a network whose links connect nodes of different types (publications and classification criteria); and (iii) the *co-occurrence link*, which is not part of the original dataset and is constructed by connecting classification criteria sharing at least one publication.

The construcution and visualization of the *network of co-occurrence* or *subject co-popularity* is analogous to the human disease network – as presented by Goh et al. 2007 – in which two disorders are connected if there is a gene that is implicated in both.[4]

Together with the analysis of each one of these networks we will present the distributions characterizing the degree, or number of links adjacent to a node.

* We'd like to thank Prof. Dr. Vinzenz Brinkmann and Dr. Ralf Biering of Stiftung Archäologie for providing the data. Dr. Martina Schwarz is credited for some useful clarifications. Special thanks go to Prof. Albert-László Barabási for making this investigation possible. Furthermore we'd like to thank the members of our audience at the AIAC 2008 BSR Poster Session in Rome for their amazing feedback.

1 Schwarz et al. 2008, the original Projekt Dyabola version, is still available via http://www.dyabola.de; Zenon, the German Archaeologic Institute's own spin off is available via http://opac.dainst.org; for Stiftung Archaeologie see http://www.stiftung-archaeologie.de.

2 For a general introduction to the science of complex networks see for example Newman Barabási Watts 2006.

3 For similar investigations using other databases in art research see Schich et al. 2008 and Schich 2009.

4 See in particular Goh et al. 2007 fig. 1.

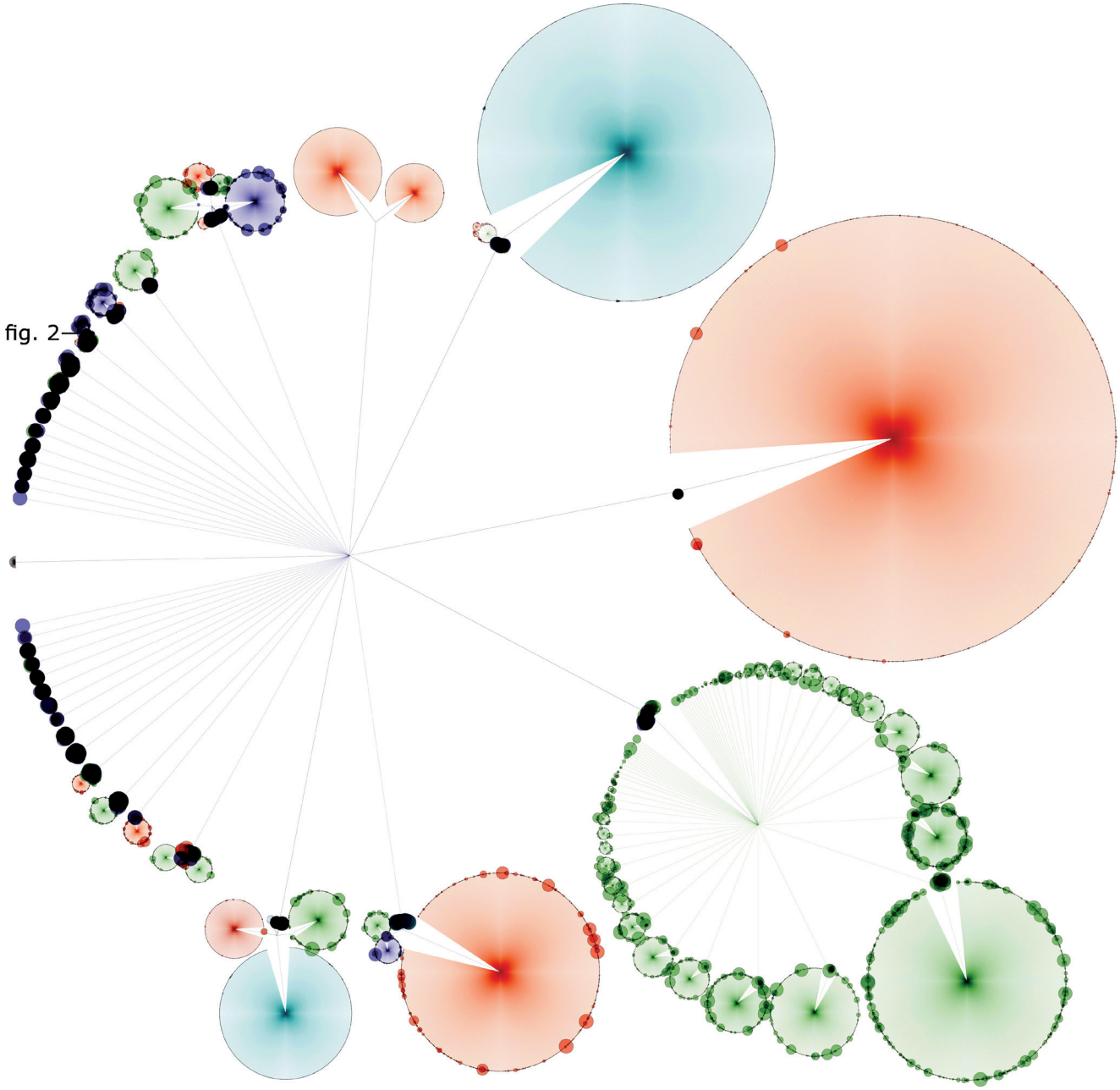

Fig. 1: The tree of subject headings as found in Archäologische Bibliographie. Classification criteria nodes, i.e. subject headings as well as keywords, are depicted as points or circles. Parent links are depicted as lines, i.e. spokes. The node size scales logarithmically with the number of bibliographic entries attached to a particular subject heading or keyword. The dataset contains locations (green), persons and institutions (red), events (turquoise) as well as subject themes (blue). There is a highly heterogenous distribution in the number of subdivisions (lines) as well as in the number of publications attached to each classification criterion (node size).

In general, we find that the distributions for each of these networks are right-skewed, a common feature of complex networks that signals that a small number of nodes in the network carries a disproportionately large number of connections. In scientific literature such distributions are often referred to as *power-law*, *long-tailed*, *heavy tailed*, *Zipf* or *Pareto distributions*. Here, we leave the issue of a precise nomenclature open, as the point we would like to make at this moment is that all the distributions we consider are approximately heavy-tailed.[5]

Results

Thematic subdivisions

Figure 1 depicts the tree of subject headings of *Archäologische Bibliographie* as of March 2008. It contains 45.924 *classification criteria*, of which 3014 are more or less predefined *subject headings* and 42.910 belong to a growing list of *keywords* forming the majority of the leaves of the tree. Every classification criterion in figure 1 is represented by a small node that is connected to a superordinate criterion via a *parent link*, represented by a line or spoke.

5 On the zoology of heterogenous distributions see Newman 2005; note that the term *long tail* was popularized by Chris Anderson in 2004 (see Anderson 2006 p. 10); however, Anderson's *long tail* contains the less connected nodes whereas in network science the tail of a distribution usually contains the hubs, due to a different assignment of the x and y axes in diagrams.

The classification criteria can be divided into a number of types, as indicated by the color of the nodes (and their respective parent link): The majority of the criteria represent locations (green), persons and institutions (red), and events and periods (turqoise), e.g. a congress in „Athens, 1962“; Subject themes (dark blue), such as „Venus“ or „Portraits of Augustus“, form only a small minority of the whole tree.
It is interesting to note that all criteria types, i.e. locations, persons, institutions, events, and subject themes, appear at multiple loci inside the hierarchical tree - some countries, for instance, are represented up to 18 times. This redundancy is due to the fact that the *tree of subject headings* is based on the card index system used since 1956 by the *German Archeologic Institute in Rome*. In this system every physical card can only be placed inside one drawer, resulting in a *strong tree*, graph-theoretically speaking, where every node can only have one parent link although synonymous cards in different drawers can be connected via an *alias link*, which is not shown in figure 1 or subject of this paper.
One of the most astonishing observations we can make in figure 1 is the highly heterogeneus size distribution of subdivisions in the tree, which we will call the *distribution of subdivisions*. It is indicated by the *node degree*, i.e. the number of parent links (spokes) pointing into a node. No matter if we pick out the whole tree, any given subbranch or a specific type of criteria, we will always find a very small number of nodes with a huge number of subdivisions and a very large number of nodes, in which the number of subdivisions fades away very quickly. Figure 5a shows the whole distribution of subdivisions. A particularly striking example for this phenonmenon is the number of sites per country in the green topography branch in figure 1, which we can see as a circle of green Pac-Man-like structures in the lower right corner. Yet another example is the number of persons in relevant keyword lists (containing researchers, ancient persons, sculptors, etc.), appearing as the red Pac-Man-like structures of deminishing size, which are distributed throughout the tree.
Zooming in, the heterogenous nature of the distribution of subdivisions appears as an ubiquitous phenomenon. Figure 2, for example, shows the *branch of plastic art and sculpture*, which contains a tiny fraction of the *tree of subject headings*. Nevertheless we find the same heterogenous distribution in the number of subdivisions in the tree. In other words, the average number of subdivisions in any part of the tree of subject headings does not characterize the system very well. Similar to other classification trees such as those found in Biology our tree is scale-free and self-similar.[6]
The growth of the distribution of subdivisions in the *tree of subject headings* depends on two factors: first, the a priori definition of drawers and partitions by the creators of the card index, and second, but more important, the local activity of all classical archeologists producing literature on specific sites or themes. In other words, the subdivisions are predefined to some extent in the form of a data model and extended by the occurence of specific classification criteria in the recorded literature.

Occurrence of themes in literature
As the occurence of new classification criteria in literature plays such an important role in the growth of the *tree of subject headings*, it is interesting to take a look at the number of times our classification criteria appear in recorded publications. Figure 5b shows the general distribution of the number of publications attached to single classification criteria in the *tree of subject headings*, showing that the heterogenous distribution of subdivisions is accompanied by another heterogenous distribution characterizing the *occurence of classification criteria* in archaeological literature.
In figures 1 and 2 the size of the nodes representing different classification criteria, as well as the font sizes of figure 2, depend logarithmically on their occurence. As a consequence, nodes, which appear twice as large as another node, occur ten times as often. Sized linearly, a popular node, like „Lysippos“ in the lower left corner in figure 2, would be comparable to the whole figure, while the smallest nodes would become invisible. Even with logarithmic sizing, the heterogenous nature of the distribution of occurence is evident. In Figure 1, especially in the corona of the larger Pac-Man-like branches, we can see large nodes within a majority of very small nodes. In Figure 2 the same phenomenon is self-evident in all subsections of the tree. No matter if we look at the distribution of occurence of classification criteria in general, among the criteria of a specific sub-branch or at the distribution of any given type of criteria, we will always find a few criteria which are super-popular and a large majority for which there is very few literature. In other words, the *distribution of occurence* of classification criteria in archaeological literature is scale-free as well as self-similar.

6 Caldarelli et al. 2004.

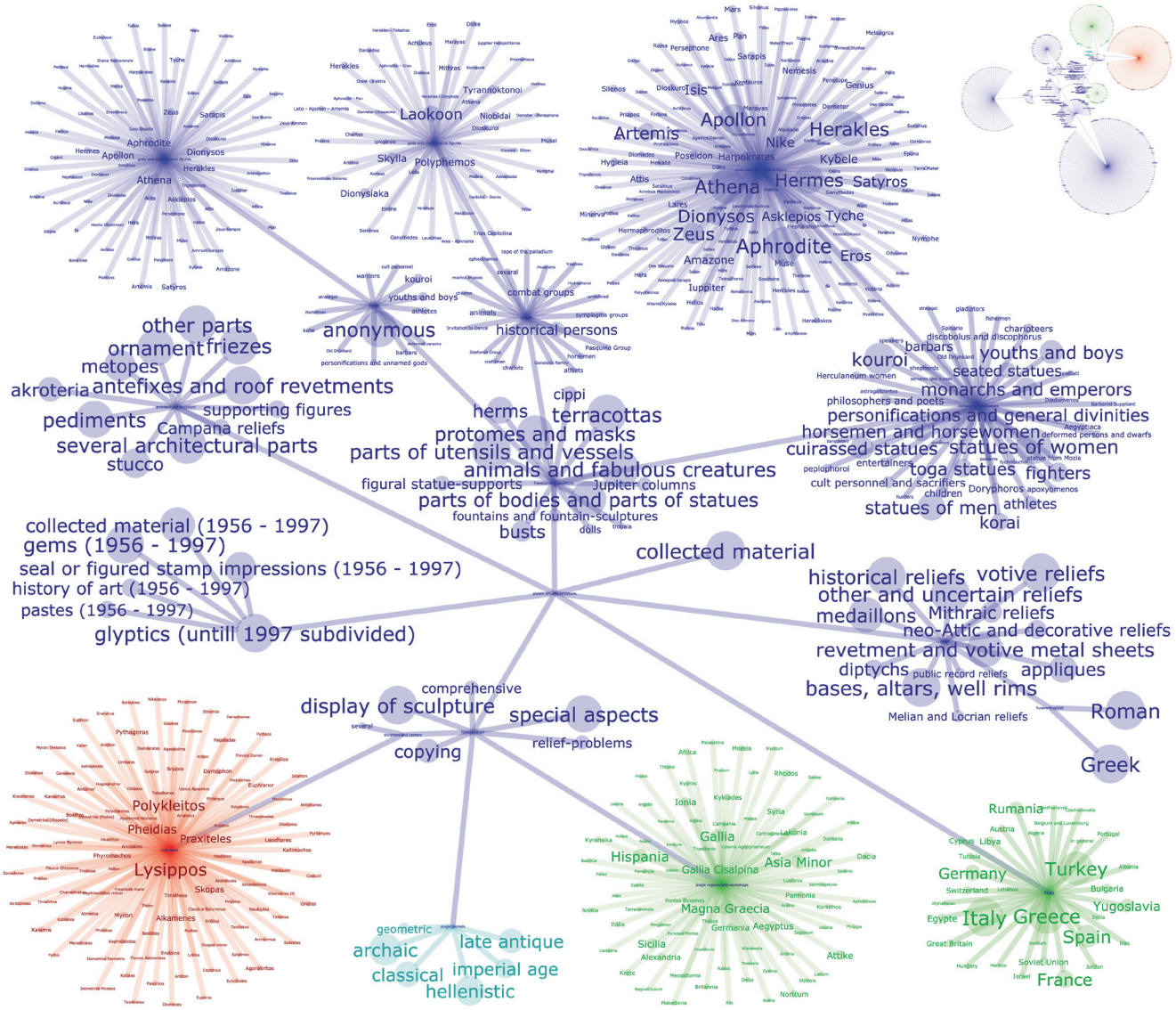

Fig. 2: The branch plastic art and sculpture, which contains a tiny fraction of the tree of subject headings. As the tree is self similar, we find the same distributions of subdivisions and number of publications attached. Note that the scaling of the node and font size is approximately logarithmic – large criteria are much more popular in reality than represented here. The small depiction in the upper right corner uses the same layout algorithm as in figure 1.

An important consequence of the self-similar nature of the *distribution of occurence* inside the *distribution of subdivisions* is the fact that we will find very popular classification criteria in a sea of very unpopular criteria at the deepest levels of the *tree of subject headings*, far removed from the casual gaze of the researcher using the tree as a browsing tool while building a specific bibliography. Therefore it would make sense to include a ranking mechanism into *Archaeologische Bibliographie* that would take into account the occurence of classification criteria as an indicator of relevance during browsing and presentation of keyword search results. (On the other hand, it has to be noted that a purely ranked keyword based search cannot replace all the benefits of subject browsing – a fact which is made clear by the emergence and growth of similar browsing structures in less aged data repositories such as Wikipedia).

Persistence of themes in literature

Extending from the question of occurence of themes in literature it is also possible to study the *persistence of themes* over the last 50 years. As a simple indication, we define persistence as the number of years in which literature on specific classification criteria occur.[7] Figure 5c clearly shows that the persistence of themes is characterized by a heavy-tailed distribution, with only a few classification criteria remaining relevant throughout the last 50 years, while most locations, persons and events are of interest in only one or a few years.

7 For a more profound investigation of persistence in a mobile phone network see Hidalgo Rodríguez-Sickert 2008.

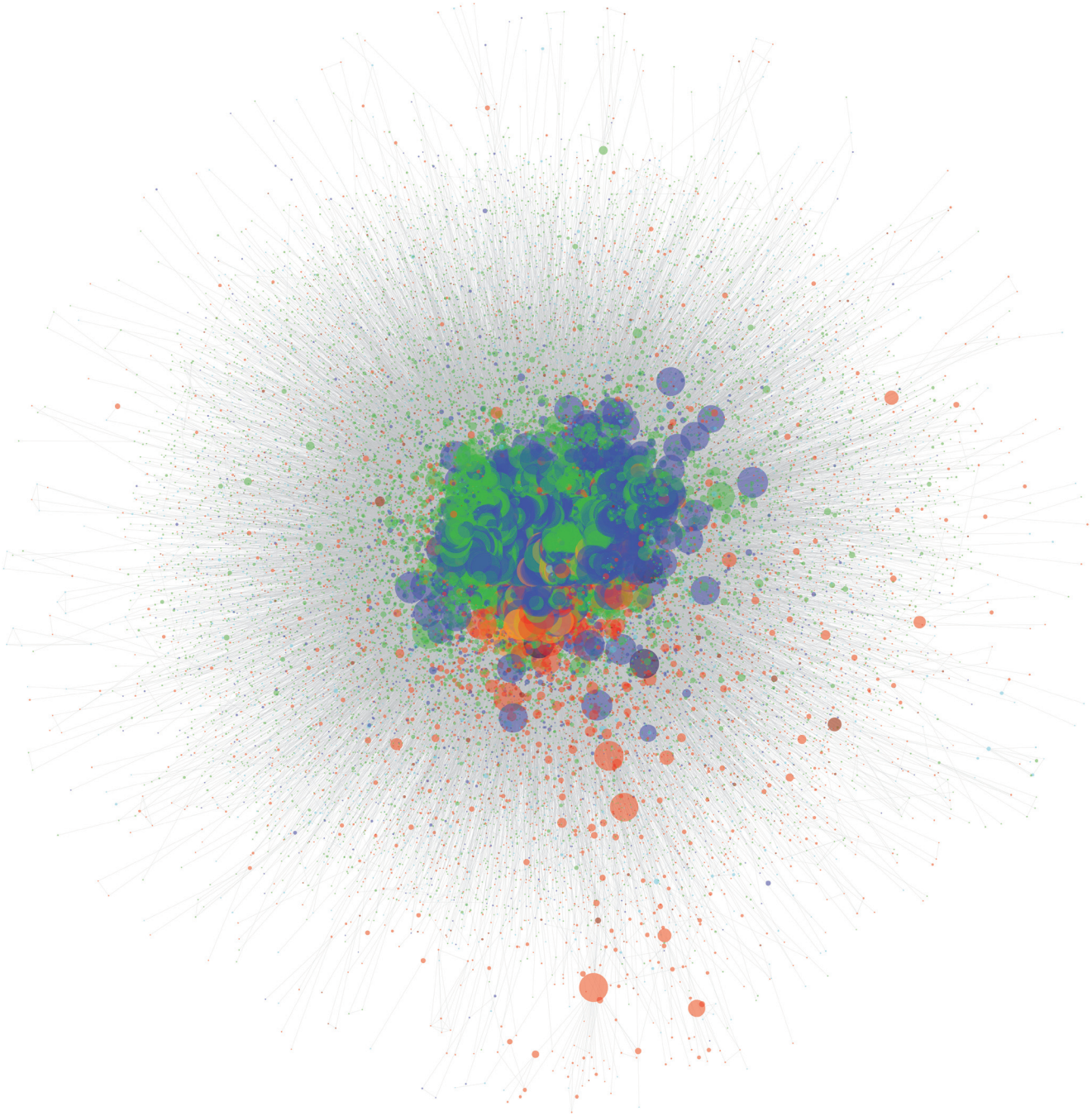

Fig. 3: The largest giant connected component (GCC) of the network of subject co-popularity. Two classification criteria are connected if they share at least one publication. The component appears as a giant hairball in which every criterion is directly or indirectly connected to all others. Note that the hairball contains a superdense core in which a few subject themes, locations, and persons form the glue that holds all other criteria together. The node size is again scaled logarithmically according to the number of publications attached to each single classification criterion – large nodes are much larger in reality than they appear.

Co-occurrence of themes in literature

We can construct a network of relations between single classification criteria, which is almost entirely based on the *local activity* of archaeologists producing the recorded literature, by connecting pairs of classification criteria that appear together in at least one publication. We assign a *weight* to each of this links equal to the number of shared publications. The resulting network of subject co-popularity for the entire classification criteria of Archäologische Bibliographie contains 29.450 nodes, which are connected by 204.056 weighted links, sharing mostly one or a few publications, except for some rare cases where up to 463 publications are shared between a pair of criteria (see the link weight distribution in figure 5e).

Figure 3 visualizes the largest, so called *giant connected component (GCC)*, which contains 95 % of the nodes and 99.6 % of all the links in the *network of subject co-popularity*. The connected component appears as a giant *hairball* in which every criterion is indirectly connected to all other criteria. As in figures 1 and 2, the size of the nodes in figure 3 is proportional to the

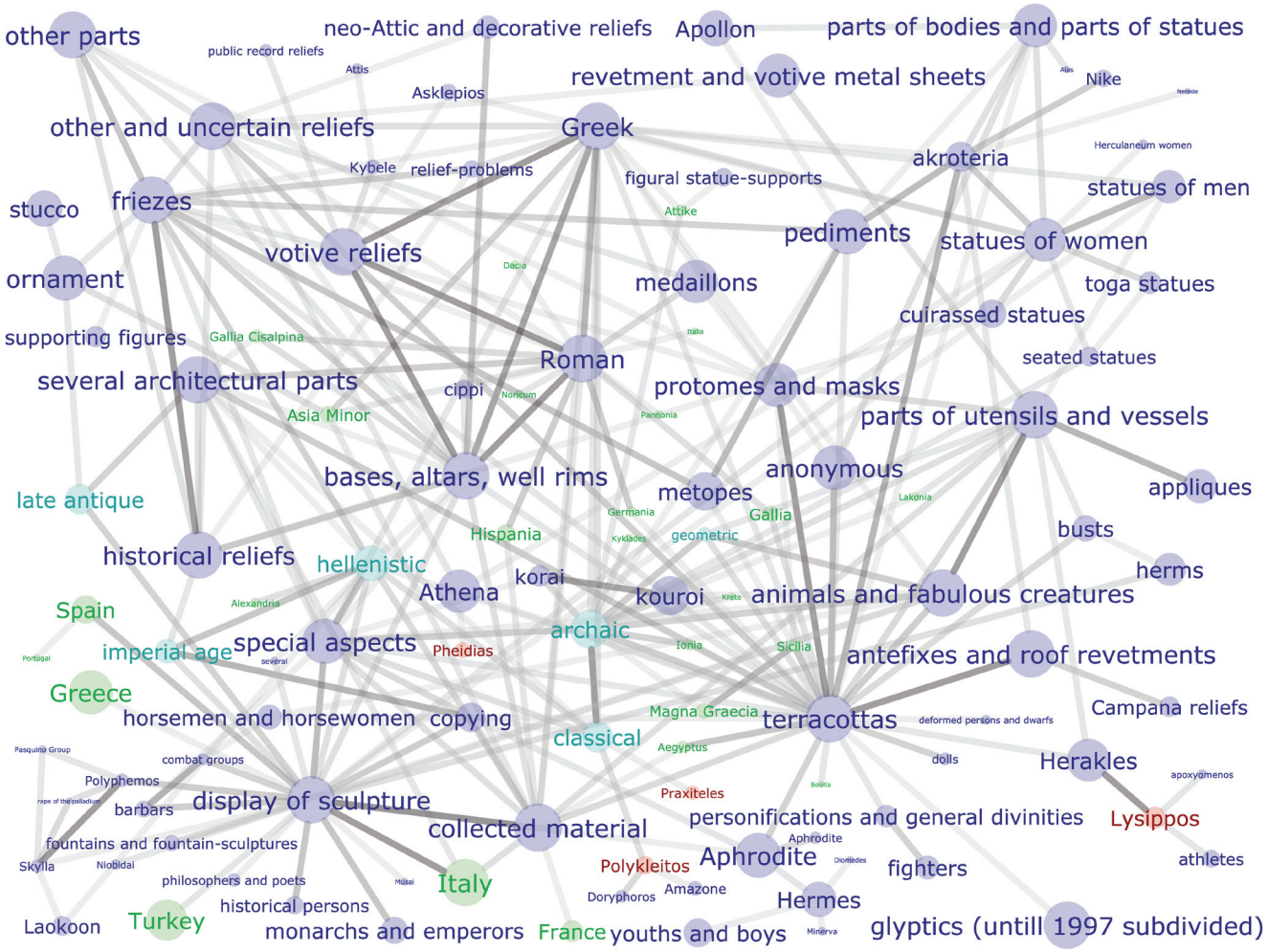

Fig. 4: The network of subject co-popularity in the branch plastic art and sculpture as depicted in figure 2. Despite a threshold of at least four shared publications the network is still superdensely connected. Note how the subject themes hold the network together and define each other by co-occurence: Hellenistic Alexandria, Classical Polykleitos, etc...

logarithm of the number of publications attached, hence in a linear scale large nodes would appear to be exponentially larger. Despite the logarithmic sizing of the nodes, the heterogenous nature of the *distribution of occurence*, inside the *network of co-popularity*, is clearly visible.

It is interesting to note that the *hairball* in figure 3 contains a superdense core in which mainly subject themes (blue) as well as a small number of popular locations (green) and persons (red) provide the glue that holds all other criteria together. This is intriguing, as we have seen that the subject themes (blue) constitute only a tiny fraction of the *tree of subject headings*. Obviously the *distribution of occurrence* (figure 5b) is closely related to the *distribution of co-occurence* (figure 5d). In other words, popular criteria are interrelated with other popular criteria in the network of subject co-popularity.

Figure 4 depicts a subsection of the *network of co-popularity* based on the *branch of plastic art and sculpture* in the *tree of subject headings* as given in Figure 2. Despite a threshold of a minimum of four shared publications in order to connect two criteria, the network is still densely connected. Almost every criterion is connected to every other criterion within a few steps. Inspecting the neighborhood of specific criteria we can observe how subject themes hold the network together and define each other by co-occurence: „Alexandria“ emerges as „Hellenistic“, „Polykleitos“ appears as „Classical“, and „display of sculpture“ is more strongly connected to „Italy“ than to „Greece“.

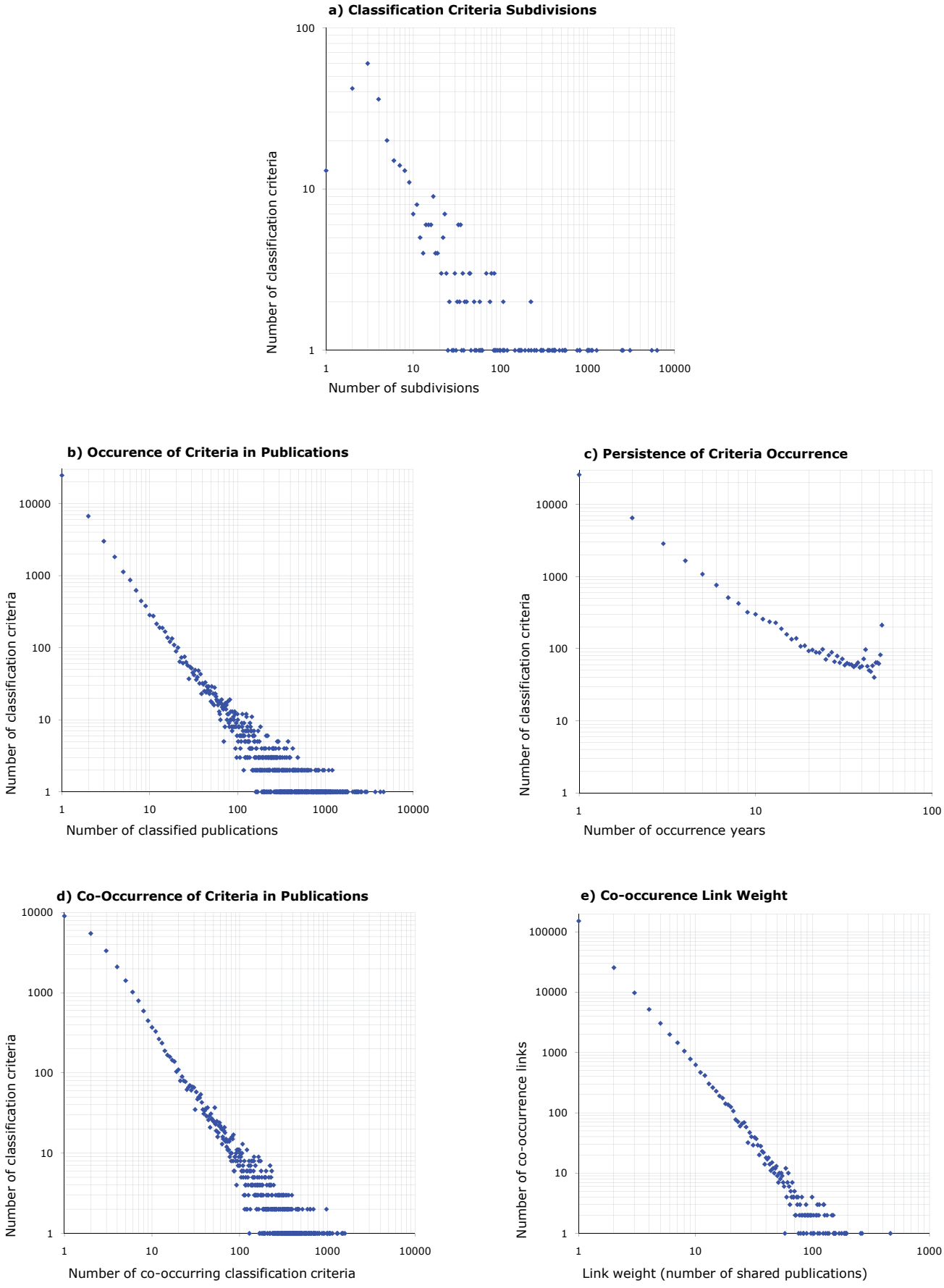

Fig. 5: a) The distribution of subdivisions, indicating the number of subdivisions for subdivided classification criteria in the tree of subject headings; b) The distribution of occurrence, indicating the number of publications classified with criteria in the tree of subject headings; c) The distribution of persistence, indicating the number of years in which the classification criteria occur; d) The distribution of co-occurrence, indicating the number of classification criteria other classification criteria are co-popular with; e) The distribution of co-occurence link weight, indicating the number of publications shared by co-popular classification criteria.

Future Work

The results provided here are a proof of concept for the fact that *Archäologische Bibliographie* contains a number of complex network properties emerging beyond the simple definition of the initial data model. Together with similar findings[8] this result is the starting point for a project at Barabasilab, funded by *German Research Foundation (DFG)*, analyzing a number of large datasets in Art Research and Archaeology. Future analysis of *Archäologische Bibliographie* will deal with the redundancy of classification criteria as well as the *bipartite nature* of the publication-classification network. In addition we plan to construct methods for breaking the *superdensely connected core of co-popularity* in order to draw a new big picture of the discipline of Classical Archeology. Our work will provide the base for an intelligent evolution of *Archälogische Bibliographie*, where each scholar would be provided with specific results according to their own research questions. The resulting methods can also be used to explore the emerging structure of other cultural heritage databases beyond their *status quo*, i.e. beyond the definition of their initial data model. Furthermore, with regards to project evaluation, this will help with future allocation of available funds.

Bibliography

Martina Schwarz et al. (2008): *Archäologische Bibliographie. The Subject Catalogue 1956 - 2008, incl. anniversary edition 50 years.* German, English, French, Italian. München: Verlag Biering & Brinkmann, Update February 2008, http://www.dyabola.de.

Mark E. J. Newman, Albert-László Barabási, und Duncan J. Watts, *The Structure and Dynamics of Networks.* Princeton: Princeton University Press 2006.

Maximilian Schich, Sune Lehmann and Juyong Park: *Dissecting the Canon: Visual Subject Co-Popularity Networks in Art Research.* in: 5th European Conference on Complex Systems, Online conference material, Jerusalem/Israel 2008, URL: http://www.jeruccs2008.org/node/114

Maximilian Schich: *Rezeption und Tradierung als komplexes Netzwerk. Der CENSUS und visuelle Dokumente zu den Thermen in Rom* (Diss. HU-Berlin Mai 2007). München: Verlag Biering & Brinkmann 2009, URN: urn:nbn:de:bsz:16-artdok-7002

Kwang-Il Goh, Michael E. Cusick, David Valle, Barton Childs, Marc Vidal, and Albert-László Barabási: *The human disease network*. PNAS, vol. 104, May 2007, pp. 8685-8690. DOI: 10.1073/pnas.0701361104

Mark E. J. Newman: *Power laws, Pareto distributions and Zipf's law.* CONTEMPORARY PHYSICS 46, no. 5 (2005): 323. DOI: 10.1080/00107510500052444

Chris Anderson: *The Long Tail.* New York: Hyperion, 2006. URL: http://www.thelongtail.com

Guido Caldarelli, Cécile Caretta Cartozo, Paolo De Los Rios, and Vito D.P. Servedio: *Widespread occurence of the inverse square distribution in social sciences and taxonomy.* PHYSICAL REVIEW E 69, 035101(R) (2004). DOI: 10.1103/PhysRevE.69.035101

César A. Hidalgo, Carlos Rodríguez-Sickert: *The Dynamics of a Mobile Phone Network.* Physica A (2008), 387(12): 3017-3024. DOI: 10.1016/j.physa.2008.01.073

Author information

Dr. des. Maximilian Schich M.A.
Affiliation: DFG Visiting Research Scientist, CCNR - BarabásiLab, Northeastern University, Boston.
Address: 110 Forsyth St., 111 Dana Research Center, 02115 Boston, MA | Tel.: +1-617-8177880 | E-Mail: maximilian@schich.info | Web: http://www.schich.info or http://www.barabasilab.com.

8 See note 3.