Exploring Co-Occurrence on a Meso and Global Level Using Network Analysis and Rule Mining

Maximilian Schich
CCNR - Northeastern University
110 Forsyth Street
Boston, MA, US 02115
+1 (617) 817-7880
maximilian@schich.info

Michele Coscia
KDDLab - University of Pisa
Largo B. Pontecorvo, 3
56125 Pisa, Italy
+39 050 2213 3136
coscia@di.unipi.it

ABSTRACT

Starting from a bipartite classification network of objects and classification criteria – in our case taken from *Archäologische Bibliographie* 1956-2007 [14] – we present a way to explore the ecology of classification co-occurrence. Enabling meso-level exploration, we construct and enrich a weighted network of classification co-occurrence with a useful lift-significance measure, based on learned association rules. Enabling global-level exploration, we use hierarchical link clustering HLC to extract sense-making communities from the co-occurrence network, taking into account that classifications can belong to multiple communities, resulting in a community overlap network. Finally, visualizing and exploring the results including evolution in time, we offer important insights regarding the structure of classical archaeology as a discipline, while making an interesting case for applying our technique to similar datasets covering other disciplines.

Categories and Subject Descriptors

E.1 **[Data]**: Data Structures - *Graphs and networks*; I.5.2 **[Computing Methodologies]**: Design Methodology - *Pattern analysis*; I.2.6 **[Computing Methodologies]**: Learning; J.5 **[Computer Applications]** - Arts and Humanities

General Terms

measurement, human factors

Keywords

archaeology, bibliography, subject classification, complex networks, co-occurrence, hierarchical link clustering, community overlap, association rules

1. INTRODUCTION

As citation indices are of limited use and literature is still not fully available in digital form, classical archaeology or the arts and humanities in general still rely more on traditional subject classificaton as other fields do. A major pain point in exploring the respective classified literature is that scholars are usually limited to relatively simple user interfaces, where they can search or query for simple lists of literature associated to sets of classifications, or hop back and forth between classifications and publications while browsing the results. In the meantime the complex ecology of classification criteria related to each other remains opaque.

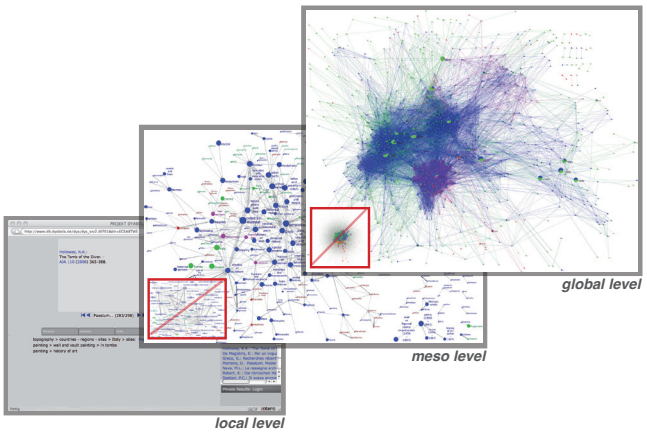

Figure 1: This paper enables meso- and global-level exploration of subject classification beyond the standard user interface of common bibliographies, improving over [13].*

Combining complex network analysis and data mining techniques in this paper, we offer a solution to this problem, enabling the exploration of a subject classification system, both on a meso- as well as on a global level, as shown in figure 1. Beyond standard user interface functionality, we are able to create a browsable set of visualizations, with which the interested scholar can explore neighboring sub-fields as well as the structure of the discipline as a whole, in a way that is more up to date and contextually superior to any written text book, as the big picture emerges algorithmically from an abundance of data that is accumulated by many actors.

As our example we use *Archäoologische Bibliographie*, i.e. a bibliographic database that collects and classifies literature in classical archaeology since 1956 [14]. Analyzing the state of 2007, our source data includes about 370.000 classified publications by circa 88.000 authors that are connected to about 45.000 classification criteria, via 670.000 classification links. Figure 2 shows a data model sketch of the database, including two additional link types which we construct within our analysis. First we generate and analyze a classification co-occurrence network from the classification link between publications and classification criteria. Second we abstract further by shortcutting from classifications to persons, resulting in an alternative perspective on classification co-occurrence in authors.

MLG'11, San Diego, CA, USA.

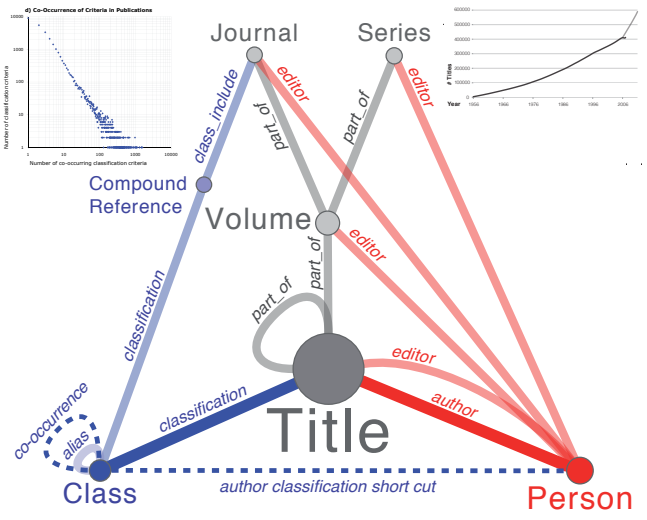

Figure 2: Data model sketch for *Archäologische Bibliographie*, including the fat-tail distribtion for classification co-occurrence in publications (upper left, see [13] for detail), and an indication of dataset growth from 1956 to 2011 (upper right).*

That our problem is not trivial becomes evident by looking at the density of the classification co-occurrence network across publications. Its giant connected component includes about 29.000 classification criteria and over 200.000 co-occurrence links, with an average diameter of 2.7. Simple node-link diagrams of co-occurrence therefore are of limited use on a meso-level, resulting in a totally useless hairball on the global level [13].

The paper is organized as follows: Section 2 indicates previous work. Section 3 details our analytical framework. Sections 4 and 5 present exemplary global as well as meso-level results respectively. Section 6 concludes the paper.

2. PREVIOUS WORK

This paper builds on previous work [13], in which Schich et al. focus on both the system of classification criteria and the bipartite network of publication–classification in *Archäologische Bibliographie*. Already discussing thematic subdivisions in the so-called *tree of subject headings*, classification occurrence frequency, co-occurrence, and persistence in literature, they bring evidence for abundant heterogeneity in the system resulting in fat-tail distributions spanning five to six orders of magnitude (see figure 2 in the upper left) – in fact legitimizing our perspective using approaches taken from the science of complex networks.

In particular our paper makes use of a method [2] taken from the area of network community finding [8, 11], combining it with a method for filtering dense networks in an intelligent way [3]. Regarding the area of data mining and learning, our paper furthermore makes use of an established technique extracting association rules [1, 9] in order to produce a sense-making lift-significance-weight in addition to regular co-occurrence. As an alternative to association rules one could also apply a weighting scheme such as TF-IDF [12], which we have avoided as larger background corpuses would have been hard to apply in our case, with classification criteria not being single ngrams, but branches of in part multilingual phrases within the strong *tree of subject headings*, where the very same term, such as a country name, can appear in multiple places within the hierarchy.

For visualization we made use of Cytoscape [15].

3. METHOD

In terms of method this paper centers on the pipline depicted in Figure 3. Starting from a given source dataset, that is a bi-partite classification network, it includes (a) a one-mode projection from object–classification to classification co-occurrence, (b) the creation and visualization of rule-mined directed lift-significance link weights in addition to regular co-occurrence weights, and (c) the creation and visualization of a link community network, using Vespignani-filtering and Hierarchical Link Clustering HLC. In our paper we use the full pipeline in Figure 3 on five source dataset snapshots as derived from *Archäologische Bibliographie*, cumulating from 1956 to each full decade until 2007. We do this for both, classification co-occurrence in publications as well as classification co-occurrence in authors – summing up to ten source dataset snapshots in total. In addition to the main pipeline in Figure 3, we run an era-discovery algorithm on the full publication dataset from 1956-2007, verifying our arbitrary decision to cumulate decade by decade. Finally we also connect communities resulting from the pipeline in Figure 3c across decades. In a more formal way the problem we solve with this pipeline can be defined as follows:

DEFINITIONS – Given a bipartite classification network of objects and classification criteria, (1) while aiming for meso-level exploration, construct a weighted network of classification co-occurrence, enriching it with a useful significance measure, which is mined using information inherent in the source network itself, and (2) while aiming for global-level exploration, algorithmically extract sense making communities from the constructed classification co-occurrence network, taking into account that classifications can belong to multiple communities, resulting in a community overlap network. Finally, given multiple snapshots of the co-occurrence network in time, (3) connect their respective community overlap network, enabling the exploration of their evolution in time.

In formal terms, our analysis starts with a set of *objects* O – i.e. in our case a set of publications or authors – and a set of associated *classification criteria* C. Elements $c \in C$ are related to objects $o \in O$ in a many-to-many fashion, meaning each classification can refer to many objects, while each object is potentially connected to many classifications. Both sets of classifications and objects grow over time. Therefore, we model our problem in the form of an *evolving unweighted bipartite graph* $G = \{O, C, S, E, T\}$, where (a) each classification c may belong to a particular *classification superclass* $s \in S$, representing the axiomatically discrete dimensions of Location, Person, Event, Period, or more general Subject Themes; (b) E is a set of triples (o, c, y), with y signifying a point in time at which the relationship between c and o has been created; (c) T is the set (c, s) which maps each classification c to its one and only one corresponding supertype s.

It is worthwhile noting that we apply our method to a single data source, while the problem definition given above is general, meaning it can also be applied to any other system that can be interpreted as a bipartite network of objects and classification criteria. Furthermore, losing only one degree of freedom, it is not mandatory that the system grows over time or supertypes are assigned to classifications.

Below the method is explained in more detail. Following data preparation (Section 3.1) we split our main analysis pipeline in two: Part one finds overlapping communities of classifications (Section 3.2), resulting in a global level abstraction of our system; Part two enriches co-occurrence with a directed lift-significance weight (Section 3.3), refining meso-level exploration. We conclude with the optional era-finding and snapshot connection (Section 3.5 & 3.6).

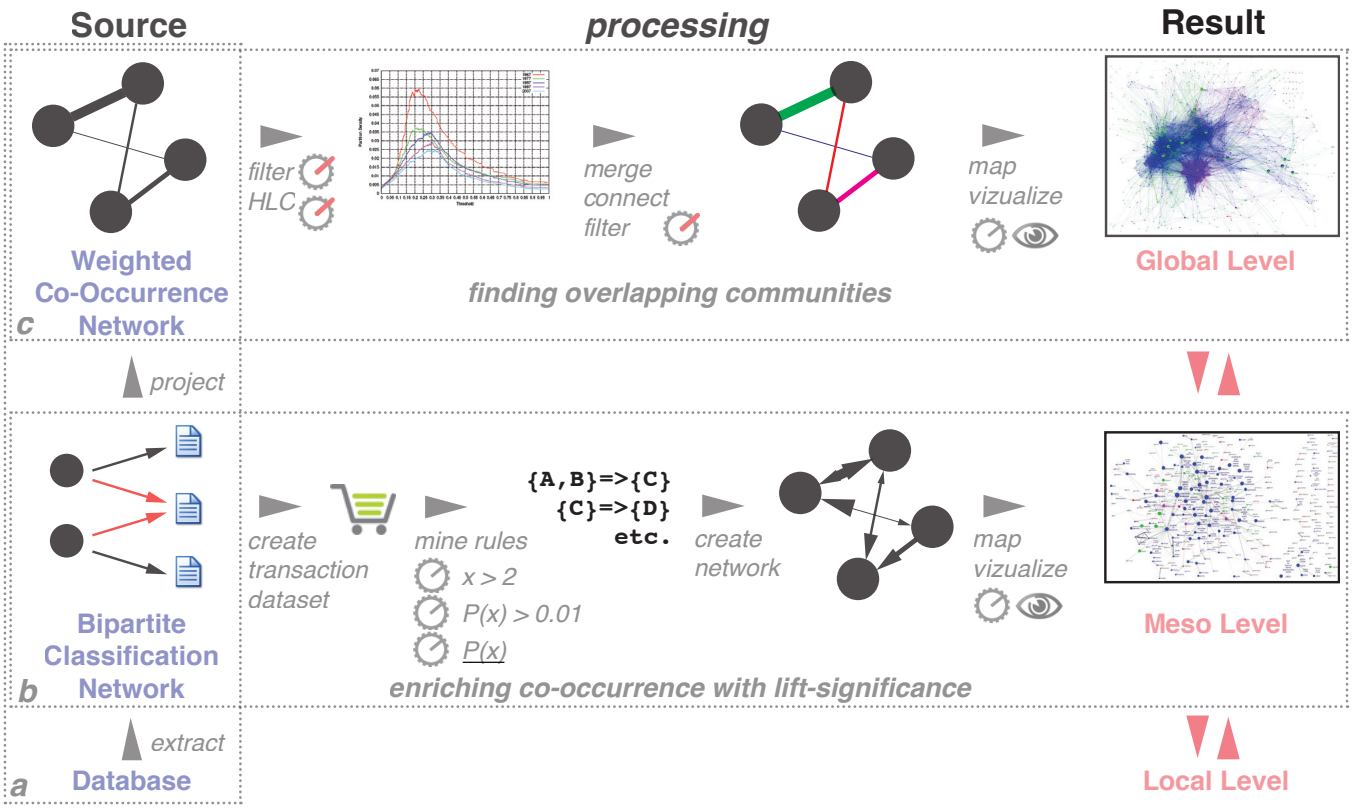

Figure 3: Data preparation, analysis, and visualization pipline as described in points 3.1 to 3.3, including (a) the one-mode projection from publication–classification or author–classification to classification co-occurrence, (b) the creation and visualization of rule-mined directed lift significance link weights in addition to regular co-occurrence weights, and (c) the creation and visualization of the overlapping community network, using Vespignani-filtering and Hierarchical Link Clustering.

3.1 Data Preparation

Regarding data preparation we follow the pipeline in figure 3a, starting from a bipartite classification network extracted from a source database, as formalized above.

For the meso-level pipeline (Section 3.3) we transform the edgeset E into a transaction dataset where each line takes the form (o; y; c1; c2; ...cn). In other words, each object o is handled as a transaction in a transactional dataset containing the list of its classifications as items, resulting in a list of adjecency lists for all $o \in O$.

For the final visualization in the meso-level pipeline and the global level pipeline (Sections 3.2 & 3.3) we project our bipartite or two-mode classification network to a one-mode network of classification co-occurrence. Projecting to the set of classifications C, here results in a weighted undirected graph G'={C;E'}, where E' is a set of triples (c1; c2; w) and w is the number of objects attached to both c1 and c2 in the original bipartite graph G.

As we are interested in co-occurrence evolution, but our implemented pipelines are not defined for evolving data, we filter our source data, producing a number of temporal snapshots, that cumulate from the beginning of the source dataset to a selected point in time. More formally, for each snapshot $d \in D$ a subgraph $Gs = \{O;C;Es\}$ will be created, in which all $(c; o; y) \in Es$ will respect the condition $y \leq d$. For d we arbitrarily choose cumulating to each full decade of our example dataset, while we also address finding optimal sets of d (in Section 3.4), and connecting multiple d (in Section 3.5).

3.2 Finding Overlapping Communities

For global level exploration we follow the pipeline in figure 3c, where we aim to provide a big picture that exposes overlapping community structure as expected to be inherent in the network of classification co-occurrence. Not enforcing classifications to belong to a single community we eventually want to build and visualize a community network, where links signify at least one shared classification.

Starting from the weighted one-mode projection of our bi-partite classification graph (Section 3.1) we want to apply an overlapping community discovery technique [7, 11] – Before we do so however, we have to deal with the extreme density of our one-mode projection, which is expected especially for bipartite classification graphs, caused by hubby objects and authoritative classification criteria. In order to get around this problem, we apply a statistical filter. Instead of a simple threshold on the edge weights, we apply a sophisticated network backbone extraction technique [3], that takes into account that in weighted networks many nodes have only low-weight connections, causing them to disappear in a naive threshold filtering. Instead of deleting all edges with a weight less than a particular value and consequently many nodes, network backbone extraction in ideal cases preserves 90% of the nodes while reducing the number of edges to 50% or lower (cf. Figure 4). To do so, for each edge (i; j) the weight is recomputed – two times for both nodes it is attached to – according to the following formula:

$$\alpha_{ij} = 1 - (k-1)\int_0^{p_{ij}} (1-x)^{k-2}\,\mathrm{d}x$$

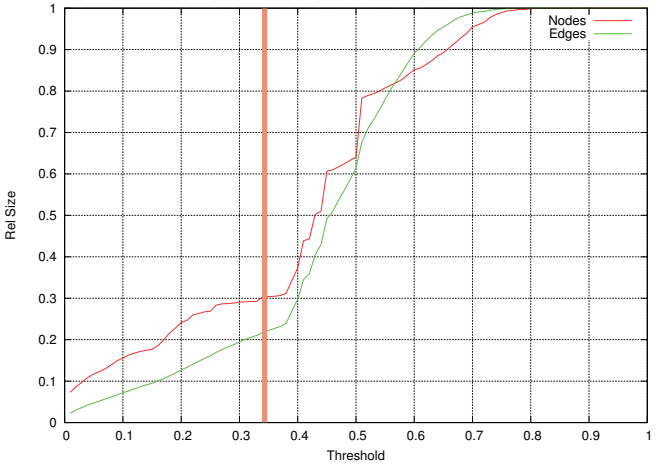

Figure 4: Relative number of nodes and edges for different values of the network backbone filtering [3] for co-occurrence in publications 1956-2007. Edges disappear quicker than nodes, making the graph sparser, as desired.

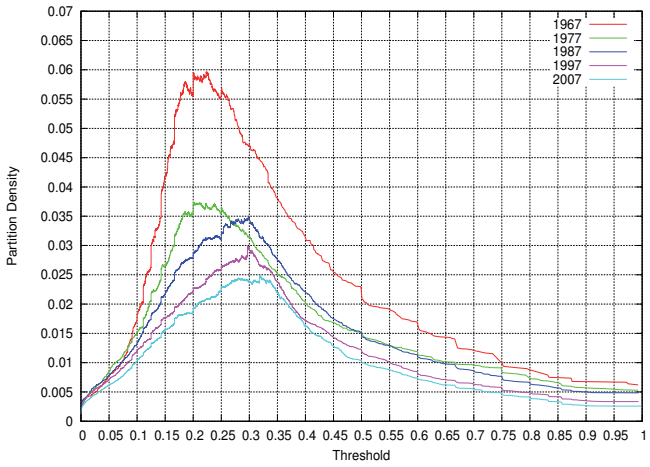

Figure 5: HLC partition density values as a function of the dendrogram cut threshold for each decade. Higher values mean denser partitions, i.e. better community division.

where k is the degree of i (or j), and p_{ij} is the normalized weight of the edge, according to the total weight of node i (or j). Those edges for which $a_{ij} \leq a$, i.e. which pass the significance test according to the threshold, are preserved in the network.

From the filtered co-occurrence network we can now extract communities. A recent approach to obtain an overlapping graph partition is to perform the community discovery on the edges instead of the nodes themselves [2,5]. From the given options we chose to apply Hierarchical Link Clustering HLC [2] as this method turned out to produce the most useful results. HLC first uncovers the hierarchical structure of the link communities in a complex network, where communities composed of a single link are recursively merged until the network itself composes one giant community. Meaningful communities are then extracted, by cutting the community dendrogram. Deciding for a meaningful cut, modularity [10] is widely used to evaluate the quality of a partition. However as this is not well defined when including overlap, plus some other drawbacks (such as the resolution limit [8]), we follow [2] evaluating the quality of each partition using the partition density D score, which is (given a partition p returning a set of link communities LC):

$$D(p) = \frac{2}{|E'|} \sum_{lc \in LC} |E'_{lc}| \frac{|E'(lc)| - (|C(lc)| - 1)}{(|C(lc)| - 2)(|C(lc)| - 1)}$$

where |E'| is the total number of edges in the network, and |C(lc)| and mlc are the numbers of nodes and edges in lc respectively. The higher D(p), the better the partition p identifies well divided clusters in the network.

Figure 5 reports the evolution of partition density for all possible dendrogram cuts in our co-occurrence network in publications for each decade (i.e. our snapshots d ∈ D). For each decade, choosing the given optimal partition p, we now obtain a set of overlapping communities LC, allowing us to produce the desired global level picture of our classification network. In order to do so, we collapse each lc ∈ LC into a single node, connecting the nodes of this network with links, whose weight is proportional to the number of nodes shared by the two communities. As each node and edge has a complex internal structure derived from the weight of the classification supertypes of all c ∈ lc, we can further enrich both nodes and edges in the visualizaton of the resulting community overlap network, by representing the nodes with pie diagrams and splitting the edges according to the inherent superclass frequency.

As the community overlap network is again very dense, and we aim for a text-book-style global picture we apply the backbone filter again [3].

3.3 Lift Significance

For meso-level exploration we follow the pipeline in figure 3b. Here we aim to visualize our simple weighted co-occurrence network of classifications c ∈ C with a more sophisticated directed significance measure. In order to do so, we perform association rule mining [1] over our transaction dataset (as introduced in Section 3.1), mining for frequent rules of co-classifications. Minimum support and confidence thresholds may be tuned depending on the phenomenon one is interested to highlight.

As a result we obtain a set R of rules in the form P(C) => c, where P(C) ∈ $P_{\geq 1}$(C), and $P_{\geq 1}$(C) is the power set of C, i.e. the set of all subsets of C, excluding Ø. Using this result, we are able to build our significance network in which the nodes are the classifications C, and the edges are triples (c1; c2; w(c1; c2)), where w(c1; c2), i.e. the significance of the relationship between c1 and c2 is defined as follows:

$$w(c_1, c_2) = \sum_{\forall r \in R. c_1 \in P(\mathcal{C}) \wedge c = c_2} \frac{supp(P(\mathcal{C}) \cup c)}{supp(P(\mathcal{C})) \times supp(c)}$$

where P(C) is the set of classifications in the left side of rule r, c is the classification in the right side of the rule r, and supp(x) is the support of the set x of classifications inside the transactional dataset. In other words, w is the sum of the lift of all rules involving c1 as one of the antecedents of the rule, and c2 is the consequence. The lift measure as such is not directed, but since we are filtering rules according to their confidence, which is directed, it follows that w(c1; c2) ≠ w(c2; c1), resulting in a directed network. This means a situation may (and does) occur, in which c1 is very significant in pointing to c2, while c2 is not so significant in pointing to c1 (see Section 5.1).

3.4 Era Discovery

Neither the global- nor meso-level pipelines above take into account time. To study evolution therefore, it is necessary to discretize the evolving source network into temporal snapshots on which the pipelines can be applied separately – raising the question, how to choose the right snapshot size?

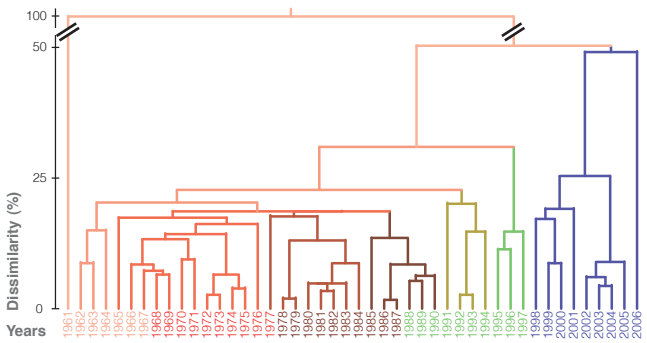

Figure 6: Era structure dendrogram of classification co-occurrence in publications of *Archäologische Bibliographie* according to [4]. Eras are colored in the tree, while our arbitrary decades are highlighted in the x-axis labels.

Looking for *eras*, i.e. periods of regular and predictable network evolution, we apply a method [4] that calculates the Jaccard coefficient (on edges and nodes) between all consecutive observations of the network, resulting in the ability to define a distance measure between groups of observations. In other words, given an observation y and the values of the Jaccard coefficient in the observations before y–1 (which is J(y−1)) and after y+1 (which is J(y+1)) the method calculates what value the Jaccard coefficient should take if y would be part of a regular era, and given its actual value J(y) how far it is actually from there:

$$dist(y) = \frac{|J(y) - (m \times y) - q|}{\sqrt{(1 + (m^2))}}$$

$$\text{where } m = \frac{J(y_{-1}) - J(y_{+1})}{y_{-1} - y_{+1}} \text{ and } q = (-y_{+1} \times m) + J(y_{+1}).$$

Using this distance measure, computed on all adjacent observations, a dendrogram is built, grouping together consecutive observations, presenting regular evolution separated from abrupt changes in trend. Figure 6 depicts the respective dendrogram for classification co-occurrence in publications of *Archäologische Bibliographie* from 1956 to 2007, with our arbitrary decades fitting surprisingly nice to the found era structure.

3.5 Snapshot Connections

Finally, given the fact that our analysis is performed in separate pipelines for each decade or snapshot, how can the snapshot results be connected? In the meso-level case the solution is trivial: All classifications are uniquely identified and can therefore be connected across snapshots. For the global level this is not true since communities are calculated for each snapshot separately. So, given community A in snapshot d and community B in snapshot d+1, can we decide if A and B are related or not – i.e. if they are equivalent, if B forked from A, or B is a merge of A and C?

In [6] the authors solve this problem with the concept of *minimum description length*, i.e. by using a data description language to produce the shortest data description possible. In our case all communities are lists of classifications, where we can calculate the relative entropy between any community pair from one snapshot to another. The relative entropy takes values from 0 (where two communities share all classifications) to +1 (where the community overlap becomes zero). Calculating the relative entropy across snapshots, we can put weighted links from a community in snapshot A to one or more communities in the subsequent snapshot B. The weight is inversely proportional to the relative entropy. Figure 10 below shows an example result.

4. GLOBAL EXPLORATION

4.1 Community Overlap

As a result of processing our source data according to the pipeline in figure 3c, we can explore the ecology of classifications in *Archäologische Bibliographie* on a global level, i.e. in form of an overlapping community network. Nodes in this network, as shown in figure 7, represent a number of classifications belonging to the respective communities, with the amount of classifications indicated by node size. Links between the communities stand for the number of classifications that are shared between them. Every classification in our system, can therefore potentially be part of multiple nodes and links in the community network. That the found configuration of communities makes sense, becomes clear while zooming into the meso-level structure of our system further below in section 5. First however, we take a look at some obvious features of the global community network.

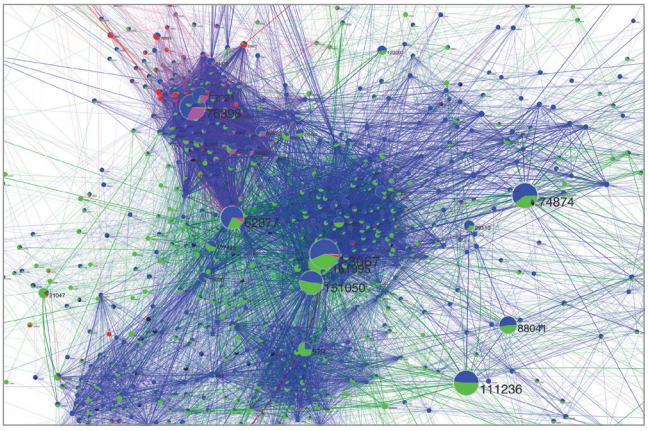

Figure 7: Detail of a community overlap network, with nodes represented as pie diagrams and edges split, indicating the inherent frequency of classification supertypes, i.e. subject themes, locations, periods, persons and objects.*

4.2 Overlap Type Distribution

One of the most striking features of the community overlap network in figure 7 is that it is NOT a hairball, but a collection of tightly connected clusters that are interconnected in a semi-tight fashion and surrounded by a sparsely connected periphery. Using superclass information in order to enrich the visualization, it becomes clear where the observed structure is rooted: Every node in our community network is depicted as a pie chart indicating the presence of classification superclasses in the respective community – blue for subject themes, green for locations, pink for periods, red for persons and black for objects and monuments. Even without knowing the detailed content of our communities it becomes immediately clear to the eye, that the superclasses are not distributed in a random way, but grouped into genres defining the tightly connected clusters.

The situation becomes even more clear if we look at the distribution of links split into their superclasses, so that say a community link containing three locations and four subject themes is split into two lines, i.e. a green line of width three, and a blue line of width four. Figure 8 shows all the location, period, and subject theme links in isolation for both co-occurrence of classifications in authors as well as publications according to the state of *Archäologische Bibliographie* in 2007. We can clearly see that subject classifications permeate throughout the whole community

network, while periods and locations co-govern certain clusters. In other words, according to *Archäologische Bibliographie*, publications and – as clusters appear to be tighter and even better defned – even more so authors in classical archaeology seem to specialize roughly on certain genres, governed by an either spatial, temporal, or a more generic conceptual perspective.

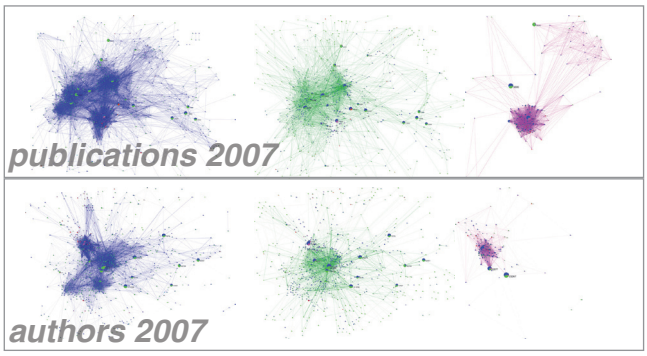

Figure 8: Links in the community overlap network corresponding to subject themes, locations, and periods are distributed in a very different way.*

4.3 Community Network Evolution

Focusing on the evolution of the community overlap network, we have applied the data processing pipeline in figure 3c five times each, cumulating classification data from 1956 for every decade from 1967 to 2007, both for classification co-occurrence in publications as well as authors. Keeping our variable threshold settings over the decades and using the same simple edge-weighted spring-embedded layout [15], we can see in figure 9 that the colored cluster structure identified in 4.1 and 4.2, comes into existence in the form of a bare skeleton of a few connected communities very early on, fleshing out to massive more differentiated proportions over the decades. The smooth development seems to legitimate our arbitrary decision to split our dataset into five decades. The fairly accurate fit of the decades to the algorithmically extracted era structure of our data in figure 6 further supports our choice. In sum we can say that the picture of community network evolution, or in other words classical archaeology according to *Archäologische Bibliographie* as a whole, does not feature large surprises – for e.g. in the form of significant phase transitions in node connectivity – but seems to grow in a smooth manner. If the smooth development reflects the evolution of classical archaeology as a discipline or is rooted in the attention towards literature on behalf of the curators of *Archäologische Bibliographie*, remains a subject of further investigation.

4.4 Community Evolution

Zooming into the evolution of communities themselves, according to the algorithm used in 3.5, reveals a more differentiated situation in detail. Looking at figure 10 for e.g. we can see two communities 27133 and 18874, which over the decade from 1987 to 1997 merge into a single community 64700, approximately averaging the fraction of associated locations, subject themes and periods, only to split up into two seperate communities 109017 and 198594 again by 2007, now concentrating periods and locations vs. periods and subject themes respectively – curiosly reflecting the often ideosyncratically perceived spat between excavation archaeologists and more art historically focused scholars spending most of their time in the library.

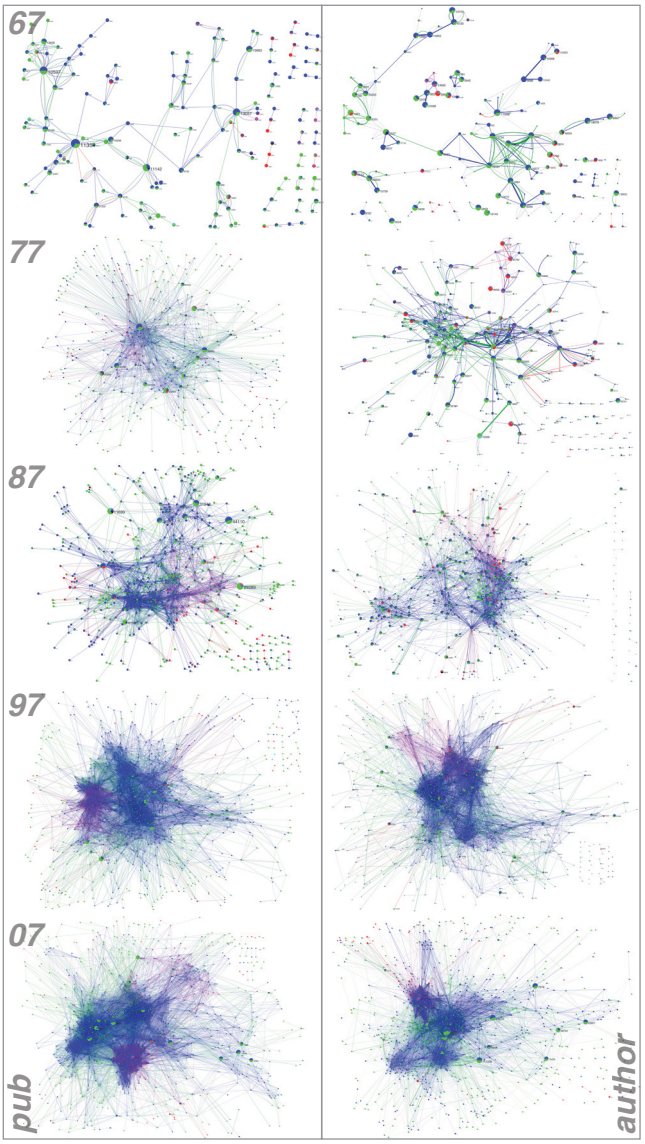

Figure 9: Both classification co-occurrence in publications as well as authors evolve over time, fleshing out structure that emerges early on in the process.*

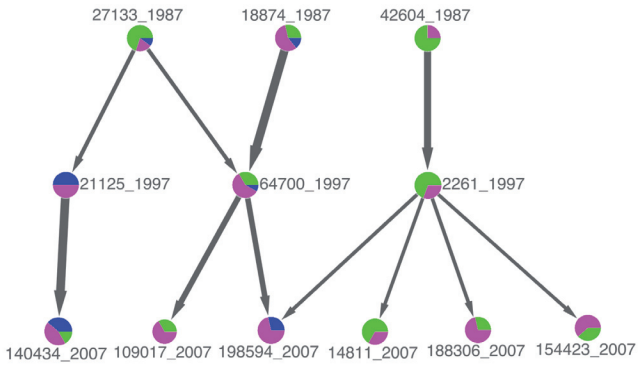

Figure 10: Communities belonging to various temporal snapshots are connected using a dedicated algorithm, revealing interesting merges and splits over time.

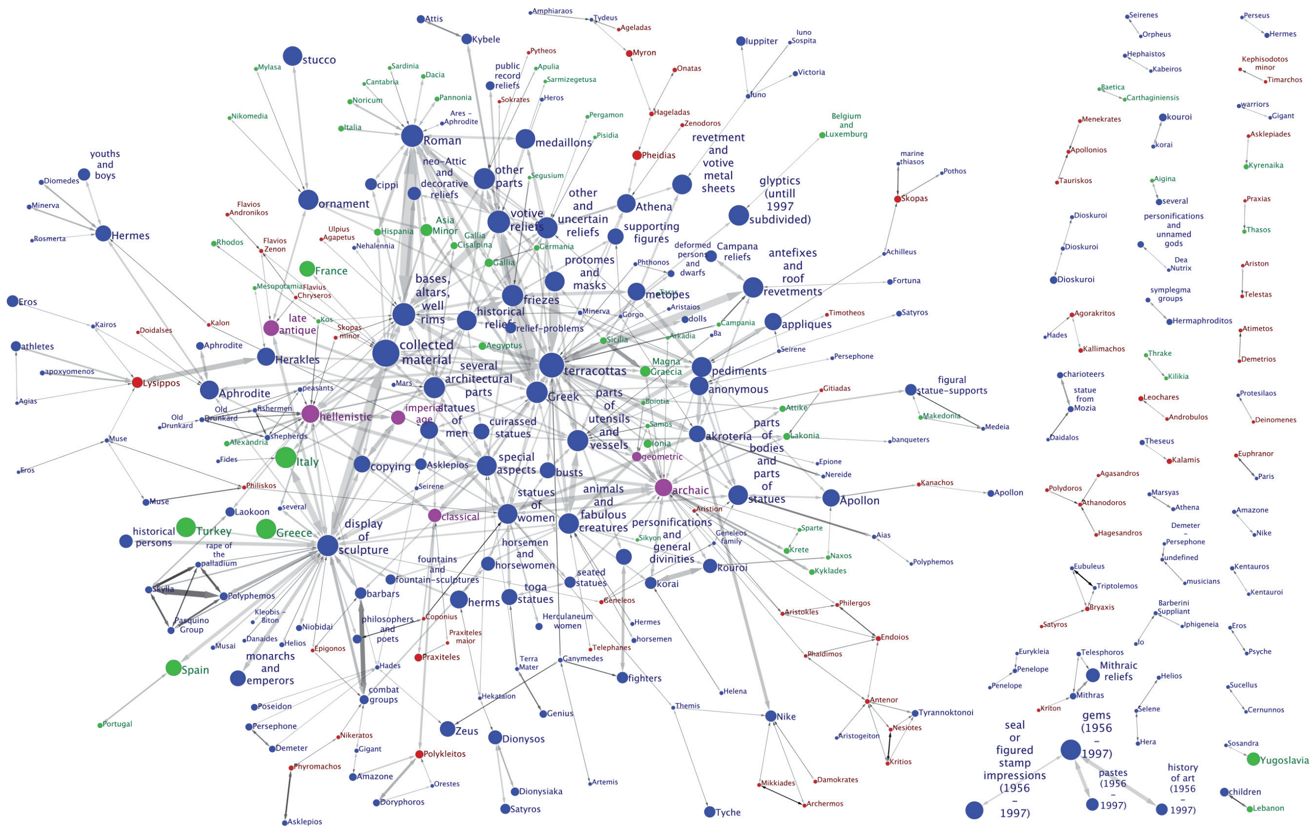

Figure 11: Classification co-occurrence in publications with lift-significance for the branch *Plastic Art and Sculpture*, i.e. a subset of classifications in the *tree of subject headings* of *Archäologische Bibliographie* in 2007. The picture, which can be seen as an instant cheat sheet for an imaginary archaeology exam, is a simple merge of two versions of the network, thresholded in different ways: Heavy co-occurrence links are taken into account if they contain at least 4 publications, equivalent to figure 4 in [13], while additional links are included if their lift significance is at least 0.056.*

5. MESO LEVEL EXPLORATION

5.1 Co-Occurrence plus Lift-Significance

As a result of the pipeline in figure 3b, we can explore the ecology of classifications in *Archäologische Bibliographie* on a meso-level, i.e. in form of a significance weighted co-occurrence network. Nodes in this network, as shown in figure 11, are the classifications themselves, with node color signifying the classification superclass – i.e. subject themes, locations, periods, persons, or objects. Node size indicates the amount of literature or number of authors associated with the classification. Links connect co-occurring classifications. Line width is proportional to a simple co-occurrence weight, i.e. the amount of literature or number of authors shared by the two connected classifications. The line color depth reflects the lift significance measure introduced in 3.3, with light grey links carrying low significance vs. darker links being highly significant. While line color depth is only a simple sum of lift significance in both directions, the respective arrow heads at both ends of the line contain information about link symmetry. This is interesting, as co-occurrence usually turns out to be symmetrical, but sometimes is remarkably directed by nature.

Figure 11 presents a striking example showing all the properties mentioned above. It depicts co-occurrence in the branch *Plastic Art and Sculpture* i.e. a subset of classifications within the *tree of subject headings* in *Archäologische Bibliographie*. As in previous work [13] we threshold the subset, taking only links into account that contain at least four publications. Improving over the previous version however, we also add highly significant links containing as few as a single publication. As a threshold for lift significance we use a rule of thumb, taking into account as many significant links as highly co-occurent ones, merging the two resulting thresholded networks to achieve the final figure.

It is interesting that the networks thresholded by heavy co-occurrence or high lift significance do not overlap much. In fact, when merged as in figure 11 they turn out to complement each other: *Greek* and *votive reliefs* for example have a very strong connection in terms of co-occurrence without high significance, which in a sense is trivial, as any archaeologist would know that both classifications are highly related. *Zeus* and *Ganymed* on the other hand share less literature, but nevertheless their connection is highly significant and should therefore be part of the picture. In fact their relation is also assymetrical, which makes sense as *Zeus*, the father of god and men, makes us think of many aspects, while *Ganymed* in sculpture is mostly depicted with *Zeus* in the form of an eagle. Taken together the networks of heavy occurrence and high lift significance result in a kind of cheat sheet for *Plastic Art and Sculpture*, where we can easily see what is often related to each other or rare and significant. Similar pictures as in figure 11 can be produced for any given branch of classifications in the *tree of subject headings*, and also, as we will see below, for more sophisticated selections of classification criteria. Before we go into detail however, let's also take a look at network evolution.

5.2 Co-Occurrence Network Evolution

As on the global level, looking at network evolution also makes sense on the meso scale. Besides the obvious growth regarding the number of classifications, and as a consequence their respective co-occurrence links, there is one particular phenomenon striking the eye in figure 12, which shows a detail of the network in figure 11 evolving from 1967 to 2007. As becomes clear over the decades, significant links tend to accumulate literature, while loosing significance. In other words as the association starts to be taken for granted the link line widens and becomes more light in color, as we can see for the links between *Nike* and *akroteria*, or *kouroi* and *korai* in figure 12. Of course, as with link symmetry, the effect shows interesting exceptions such as the highly significant clique of *Polyphemos*, *Skylla*, *Pasquino Group*, and *rape of the palladium* that we can spot on the left side periphery in figure 11. Given the spectacular uniqueness of the sculptures in question and the related controversial discussion in the literature, it is not a surprise that the associated links stayed significant over four decades while accumulating more and more literature.

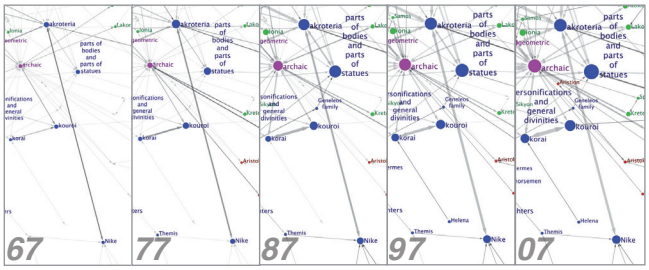

Figure 12: Classification co-occurrence evolution clearly shows that initially highly significant, i.e. dark links become less significant and wider as they accumulate literature.*

5.3 Mutual Class Self-Definition

Another interesting phenomenon on the meso level is the mutual self-definition of classifications across co-occurrence links. In previous work [13] we have already mentioned some striking examples for *Plastic Art and Sculpture* regarding this effect. Here we present another example that highlights the inherent potential: For Figure 13 we chose all classifications in the branch *Named Portraits* (across publications in 2007), thresholding both co-occurrence ≥ 2 and lift-significance ≥ 0.06 in a minimal way. Again the figure, which only shows the largest connected component of the result, can be used as a cheat sheet, indicating the relations of portraits from *Augustus*, to *Phillippus Arabs* at the end of the Roman empire, with lift significance highlighting relations between strongly connected types such as *Caracalla*, *Septimius Severus* and *Geta*. In general terms this means our approach provides easy access to highly specialized fields that are hard to explore using a regular user interface that browses bibliographic classifications

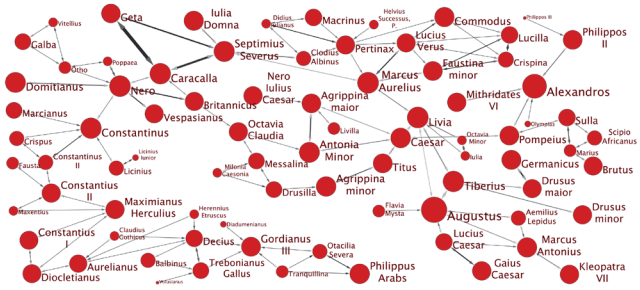

Figure 13: Mutual self-definition of *Named Portraits*.*

on a local level. As similar insights can easily be produced for all areas covered by *Archäologische Bibliographie*, the respective visualizations call for being used to complement classic textbook introductions to classical archaeology.

5.4 Ego-Networks vs. Communities

An alternative starting point in exploring the ecology of classifications in our system – beyond picking predefined branches of the *tree of subject headings* – is to begin with a single classification of interest. Here, a seemingly obvious approach would be to draw the ego-network, meaning the network of all links between classifications, the classification of interest is related to – in equivalence to the widespread basic diagrams of friendships between our own friends in popular social network platforms.

Unfortunately the ego-network strategy does not work for our co-occurrence network, as the average network diameter is only 2.7, making it very likely that the result contains an almost fully connected clique. An excellent example is the ego-network of *Paestum* – an important and popular archaeological site in Italy. Even worse than raw, thresholding has almost no effect on this structure: In fact if we threshold heavily for co-occurrence ≥ 25 – while lift significance is virtually irrelevant – the picture starts to get clearer, but we only isolate what could be called the generic Italian core of classical archaeology, where *Paestum*, even though popular, only appears in the very periphery of a large cluster, connected to a few even more peripheral events, and the obvious fact that it is known for *temples*.

The solution to the problem of dense ego-networks is to harness our global level community overlap network, from which we can pick all communities in which *Paestum* appears. Looking into those communities on a meso level it turns out, we can learn in a very straight forward way what *Paestum* is really about. Figure 14 shows the relevant section of the global community overlap network, surrounded with the meso level co-occurrence networks for the respective communities. We can see that the community size distribution is heterogeneous. Let's look into some of them: Community 6696 already improves over the basic ego-network, as it embeds *Paestum* into the core of classic archaeology including relevant classifications that are more than one hop away. The smallest communities such as 68054 tell us that *Paestum* is about *temple*, *capitals*, *planning orders*, *building construction*, similar to a couple of strikingly related sites. Community 144461 dates *Paestum* to the *Greek* period, again as a striking example for *temples*. Community 78265 provides a hint that *architectural parts* from *Paestum* were reused later in Roman buildings such as the *Palatine* palaces. Community 137152 finally provides a wider context of *Paestum* including *tombs*, implicitly pointing to literature regarding the famous *tomb of the diver* among others – in sum a pretty accurate description of what Paestum is about, accessible in an easy way, even to the non-specialist.

6. CONCLUSION

Summing up, we have presented a way to explore a complex system of subject classification co-occurrence, by combining network filtering, community finding and association rule mining. As a result we can now explore *Archäologische Bibliographie* on three levels. To the standard local level user interface we have added a meso-level network of significance-weighted co-occurrence that allows us to explore the regional neighborhood of individual (groups of) classifications. Furthermore we also provide a global level community overlap network, that allows us to grasp the big picture of classical archaeology in an intuitive way.

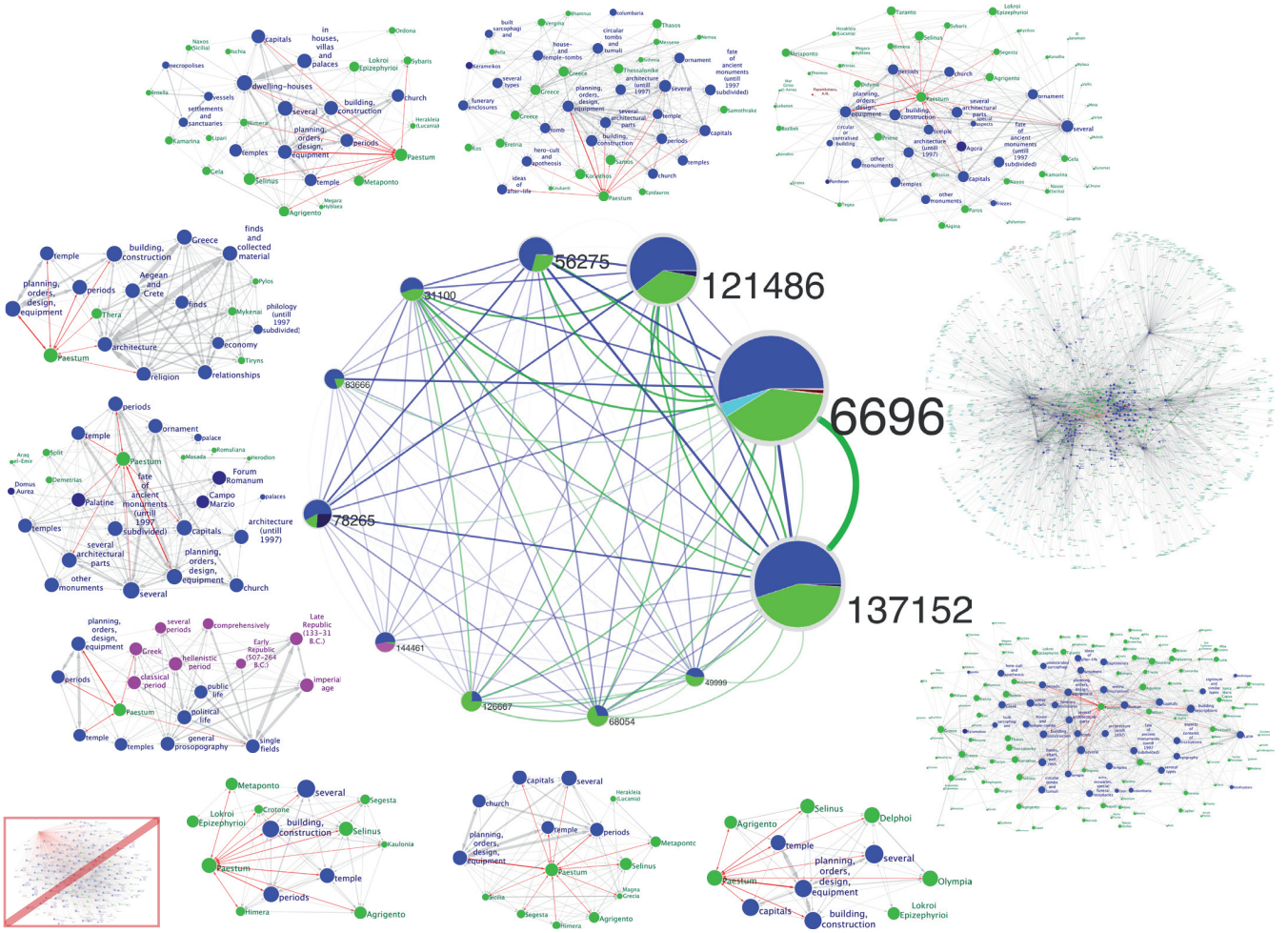

Figure 14: Combining global and meso-level exploration by zooming into overlapping communities containing a given classification – here *Paestum* – reveals its meaning even to the uneducated eye, improving significantly over simple ego-networks (see 5.4).*

7. REFERENCES

Acknowledgements: Maximilian Schich is a German Research Foundation (DFG) fellow, hosted by Albert-László Barabási; Michele Coscia is a recipient of the Google Europe Fellowship in Social Computing, and this research is supported in part by this Google Fellowship.

[1] R. Agrawal, T. Imielinski, A. Swami, Mining Association Rules Between Sets of Items in Large Databases, SIGMOD (1993).

[2] Y.-Y. Ahn, James P. Bagrow, S. Lehmann, Link communities reveal multiscale complexity in networks, Nature 466, 761-764 (2010).

[3] M. Angeles Serrano, M. Boguna, A. Vespignani, Extracting the multiscale backbone of complex weighted networks, PNAS vol. 106 no. 16 6483-6488 (2009).

[4] M. Berlingerio, M. Coscia, F. Giannotti, A. Monreale, D. Pedreschi, As Time Goes By: Discovering Eras in Evolving Social Networks, PAKDD (2010).

[5] T. Evans, R. Lambiotte, Line Graphs, Link Partitions and Overlapping Communities, Phys. Rev. E 80 016105 (2009).

[6] J. Ferlez, C. Faloutsos, J. Leskovec, D. Mladenic, M. Grobelnik, Monitoring Network Evolution using MDL, ICDE (2008).

[7] S. Fortunato, Community detection in graphs, Physics Reports 486, 75-174 (2010).

[8] S. Fortunato, M. Barthelemy, Resolution limit in community detection, PNAS vol. 104(1): 36-41 (2006).

[9] J. Hipp, U. Güntzer, G. Nakhaeizadeh, Algorithms for association rule mining – a general survey and comparison, ACM SIGKDD Explorations Newsletter, Volume 2 Issue 1, 2000

[10] M. E. J. Newman, Modularity and community structure in networks, PNAS vol. 103, 8577-8582 (2006).

[11] M. A. Porter, J.-P. Onnela, and P. J. Mucha, Communities in networks, Notices Of The American Mathematical Society, vol. 56, p. 1082, 2009.

[12] T. Roelleke, J. Wang, TF-IDF uncovered: a study of theories and probabilities, Proc. of the 31st ann. int. ACM SIGIR conf. on R&D in information retrieval, 2008.

[13] M. Schich, C. Hidalgo, S. Lehmann, J. Park, The Network of Subject Co-Popularity in Classical Archaeology, Bolletino di Archaeologia On-line (2009). urn:nbn:de:bsz:16-artdok-7151

[14] M. Schwarz et al., Archäologische Bibliographie. Online-Database. Munich: Biering & Brinkmann, 1956-2011 URL: http://www.dyabola.de (Update February 2008)

[15] P. Shannon, A. Markiel, O. Ozier, N.S. Baliga, J.T. Wang, D. Ramage, N. Amin, B. Schwikowski, T. Ideker, Cytoscape: a software environment for integrated models of biomolecular interaction networks, Genome Res. 2003 Nov;13(11):2498-504.