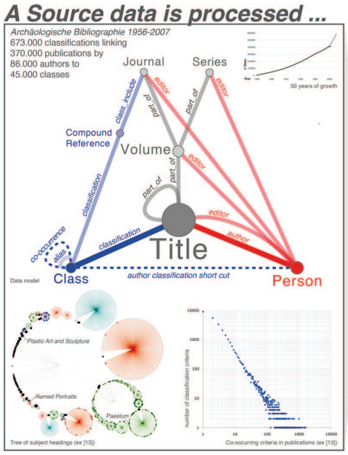
Exploring Co-Occurrence on a Meso and Global Level Using Network Analysis and Rule Mining

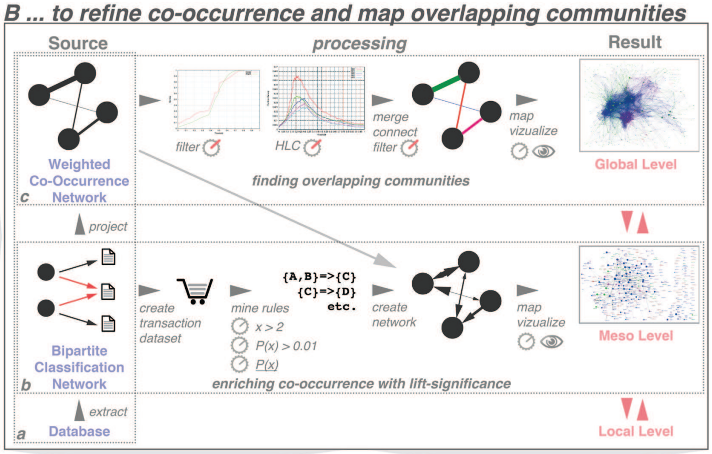
Maximilian Schich
CCNR, Northeastern University
110 Forsyth Street, Boston MA 02115, USA
+1 (617) 817-7880 – maximilian@schich.info

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

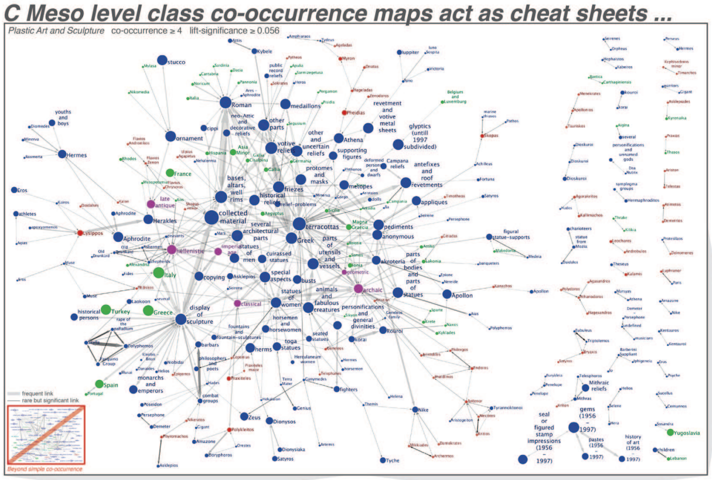
A Source data is processed ...


Archäologische Bibliographie 1956-2007
673.000 classifications linking
370.000 publications by
86.000 authors to
45.000 classes
Journal
Series
50 years of growth
class_include
editor
part_of
Compound Reference
Volume
classification
co-occurrence
alias
Title
author
Class
Person
author classification short cut
Data model
Plastic Art and Sculpture
Named Portraits
Paestum
Tree of subject headings (ex [13])
number of classification criteria
Co-occurring criteria in publications (ex [13])

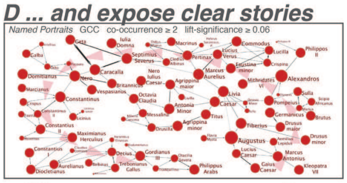
B ... to refine co-occurrence and map overlapping communities



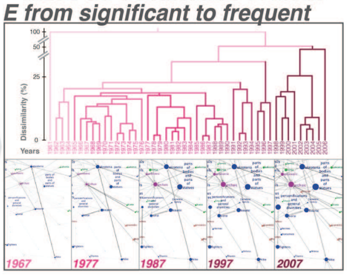
C Meso level class co-occurrence maps act as cheat sheets ...



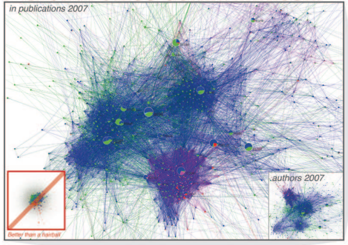
D ... and expose clear stories



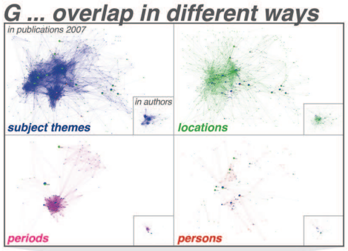
E Meso co-occurrence evolves from significant to frequent



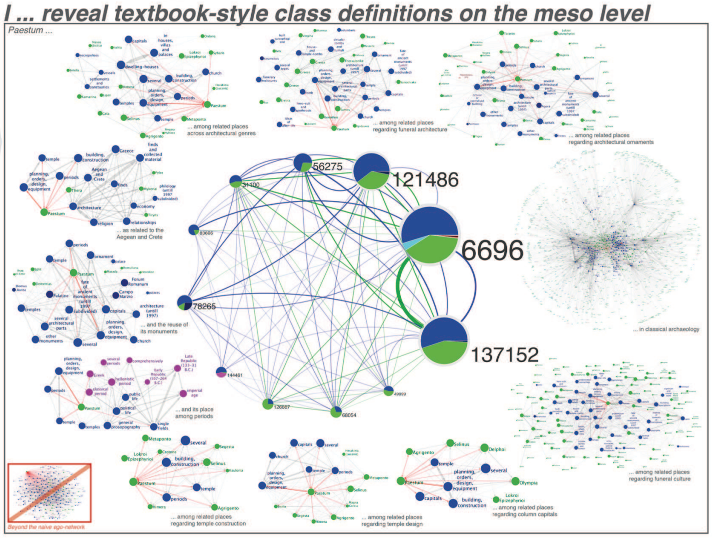
F Global class communities...



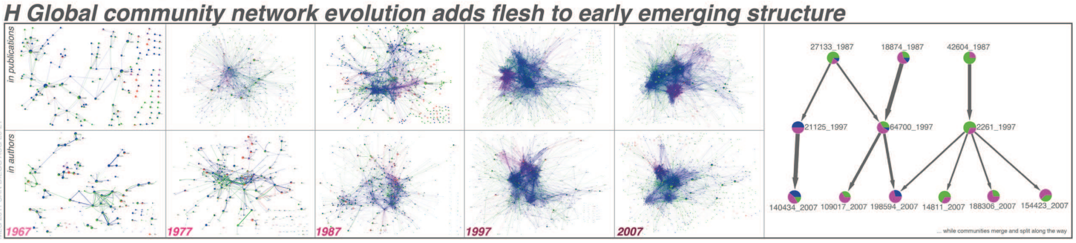
G ... overlap in different ways



I ... reveal textbook-style class definitions on the meso level


Paestum ...
... among related places across architectural genres
... among related places regarding funeral architecture
... among related places regarding architectural ornaments
... as related to the Aegean and Crete
56275
121486
6696
137152
78265
... and the reuse of its monuments
... and its place among periods
... in classical archaeology
... among related places regarding funeral culture
... among related places regarding temple construction
... among related places regarding temple design
... among related places regarding column capitals
Beyond the naive ego-network

H Global community network evolution adds flesh to early emerging structure



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.

Walk through the figures

The figure sequence makes clear how we enable meso- and global-level exploration of subject classification beyond the standard user interface of common bibliographies:

A – A data model sketch for our source dataset *Archäologische Bibliographie*, including **(below left)** a visualization of the inherent *tree of subject headings* with general subject themes, locations, persons, periods, and events; **(below right)** the fat-tail distribtion for classification co-occurrence in publications (both taken from [13]); **(up right)** an indication of dataset growth 1956-2007.

B – Our data analysis and visualization pipeline includes **(a)** a one-mode projection of the publication–classification or author–classification networks to classification co-occurrence, **(b)** the creation and visualization of *rule-mined directed lift-significance weights* in addition to the regular *co-occurrence link weight*, and **(c)** the creation and visualization of an overlapping community network, using Vespignani backbone filtering and Hierarchical Link Clustering (HLC). The **(inset plots)** in pipeline c show a backbone filtering phase transition, where edges disappear quicker than nodes, and HLC partition density distributions for five decades 1967-2007, both of which provide meaningful thresholds for further processing.

C – An example for *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*. The picture, which can be seen as an instant cheat sheet for an imaginary archaeology exam, is a simple merge of two identical networks, thresholded in different ways: Heavy co-occurring links are taken into account if they contain at least 4 publications, while additional links are included if their lift-significance is at least 0.056.

D – Another example, equivalent to figure C, where *Named Portraits* mutually define themselves and connect to a coherent story from the Roman republic to the end of the empire.

E – **(above)** An era structure dendrogram of *classification co-occurrence in publications*, where algorithmically computed eras are colored in the tree, while our arbitrarily chosen decades are highlighted in the x-axis labels. **(below)** Classification co-occurrence evolution includes initially highly significant, i.e. dark, links that become less significant and wider as they accumulate literature over five decades. Exceptions point to controversial or special topic clusters, such as the 4-clique around *Skylla*, *Polyphemos*, *Pasquino group* and *rape of the palladium* in figure C that stays highly significant over several decades.

F – Details of the global *class community overlap network* for co-occurrence in publications and authors, with nodes represented as pie diagrams and edges split by color, indicating the inherent frequency of classification supertypes, i.e. subject themes, locations, periods, persons and objects.

G – Isolating links in the community overlap network F by color, corresponding to subject themes, locations, periods, and persons, reveals that link supertypes are distributed in very different ways.

H – **(left)** Both classification co-occurrence in publications as well as in authors evolve over time by fleshing out structure that emerges early on in the process. **(right)** Communities belonging to various temporal snapshots, can be connected using a dedicated algorithm that reveals interesting merges and splits over time, indicating both diversification and specialization.

I – Combining global and meso-level exploration by zooming into overlapping communities containing a given classification – here the Italian site *Paestum* – uncovers its meaning even to the uneducated eye, clearly improving over a simple ego-networks that often present themselves as a massive featureless clique.

Conclusion

In this poster we have shown that subject themes in classical archaeology, as recorded in *Archäologische Bibliographie*, are granular components of a complex system, which we can explore both on a meso level (where themes are connected by co-occurrence) as well as on a global level (where theme communities are connected by theme overlap).

Figures F and G point to another, even more global level, as we can spot obvious clusters with our bare eyes. In order to extract these conceptual continents of the academic discipline, which might be more manifold in other datasets, it seems natural to run our pipeline c for another time.

Similar visualizations and browsable sets, which can be explored by the respective scholars, are produceable for any given library classification such as taken from arXiv, OCLC, Europeana, and maybe in not too far a future even from Google Books. As such, the poster exemplifies the usefulness of complex systems approaches, enabling a wider audience to explore and understand the complexity they are exposed to every day.

Reference summary

[1] Agrawal Imielinski Swami SIGMOD1993 (Mining Association Rules ...)
[2] Ahn Bagrow Lehmann NATURE 2010 (Link Communities Reveal Multiscale Complexity...)
[3] Angeles-Serrano Boguna Vespignani PNAS 2009 (Extracting the multiscale backbone ...)
[4] Berlingerio Coscia Giannotti Monreale Pedreschi PAKDD 2010 (As Time Goes By ...)
[5] Evans Lambiotte PRE 2009 (... and Overlapping Communities ...)
[6] Ferlez Faloutsos Leskovec Mladenic Grobelnik MDL ICDE 2008 (... Network Evolution ...)
[7] Fortunato PHYSICS REPORTS 2010 (Community detection in graphs)
[8] Fortunato Barthelemy PNAS 2006 (Resolution limit in community detection)
[9] Hipp Güntzer Nakhaeizadeh ACM SIGKDD 2000 (Algorithms for association rule mining)
[10] Newman PNAS 2006 (Modularity and community structure in networks)
[11] Porter Onnela Mucha NOTICES OF THE AMS 2009 (Communities in networks)
[12] Roelleke Wang ACM SIGIR 2008 (TF-IDF uncovered)
[13] Schich Hidalgo Lehmann Park BARCH ON-LINE 2009 (... Subject Co-Popularity ...)
[14] Schwarz et al., Archäologische Bibliographie. Online-Database. Munich: Biering & Brinkmann, 1956-2011 URL: http://www.dyabola.de (Update February 2008)
[15] Shannon et al. GENOME RES. 2003 (Cytoscape)

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.

* This work underwent peer-review in multiple related disciplines. Besides being accepted for publication at KDD MLG'11, San Diego, CA, USA, it has also been presented to an audience of archaeologists at CAA 2011 in Beijing, China, and will be shown as a poster to the complex systems crowd at ECCS 2011 in Vienna, Austria. Official publication URL: http://www.cs.purdue.edu/mlg2011/papers/paper_22.pdf

Full paper: http://goo.gl/Osrl5