Ontology decomposition process based on structural dependencies among concepts

Alena Gregová

TU Liberec

Seminar KEG March 25 2010

Contents

Introduction

- An increased interest on modularization
  - Obtain the necessary knowledge
- Reuse, scalability, maintenance
- The increasing awareness of the benefits of ontologies in open and weakly structured environments – creation of ontologies for real world domains – complex domains (medicine) contain thousands of concepts – new issues

New issues

- Maintenance
  - Large onotologies cannot be created and maintained by a single person
  - Requires team of experts from different organizations
- Publication
  - Large ontologies are created to provide a standard model of the domain
  - Interest on a specific part of the overall domain
- Validation
  - The nature ontologies require a high degree of quality of the respective model
  - Validation by different experts – large ontologies – difficult to understand
- Processing
  - On a technical level – large ontologies – scalability problems

A definition of modularization

- Allows to understand a large ontology as a set of smaller parts – modules – the decomposition process
- Another view – composition process – connection of smaller parts to a larger ontology

Goals of modularization

- Scalability – two views
  - Scalability for a search knowledge
  - Scalability for an evolution and maintenance
- Understandability
  - Size of ontologies
  - Users of ontologies – human or an intelligent agent
  - Presentation form
- Reuse
  - Reuse of already generated modules

A definition of module

- **Module**
  - reusable component, which is self-contained, bears a relationship to other modules
  - Is self-contained without references to other concepts
  - As an object representing minimum set of axioms, which makes sense
  - $Mi(O)$ – a set of axioms, $Sig(Mi(O)) \subseteq Sig(O)$
    - Partition of ontology to set of modules $\{M_1,\ldots,M_k\}$
  - $O = (C,R) \rightarrow O_M = (C_M, R_M)$

$$C_M \neq \emptyset \wedge C_M \subseteq C$$

$$R_M \subseteq R$$

$$O_M \subseteq O$$

Intuition

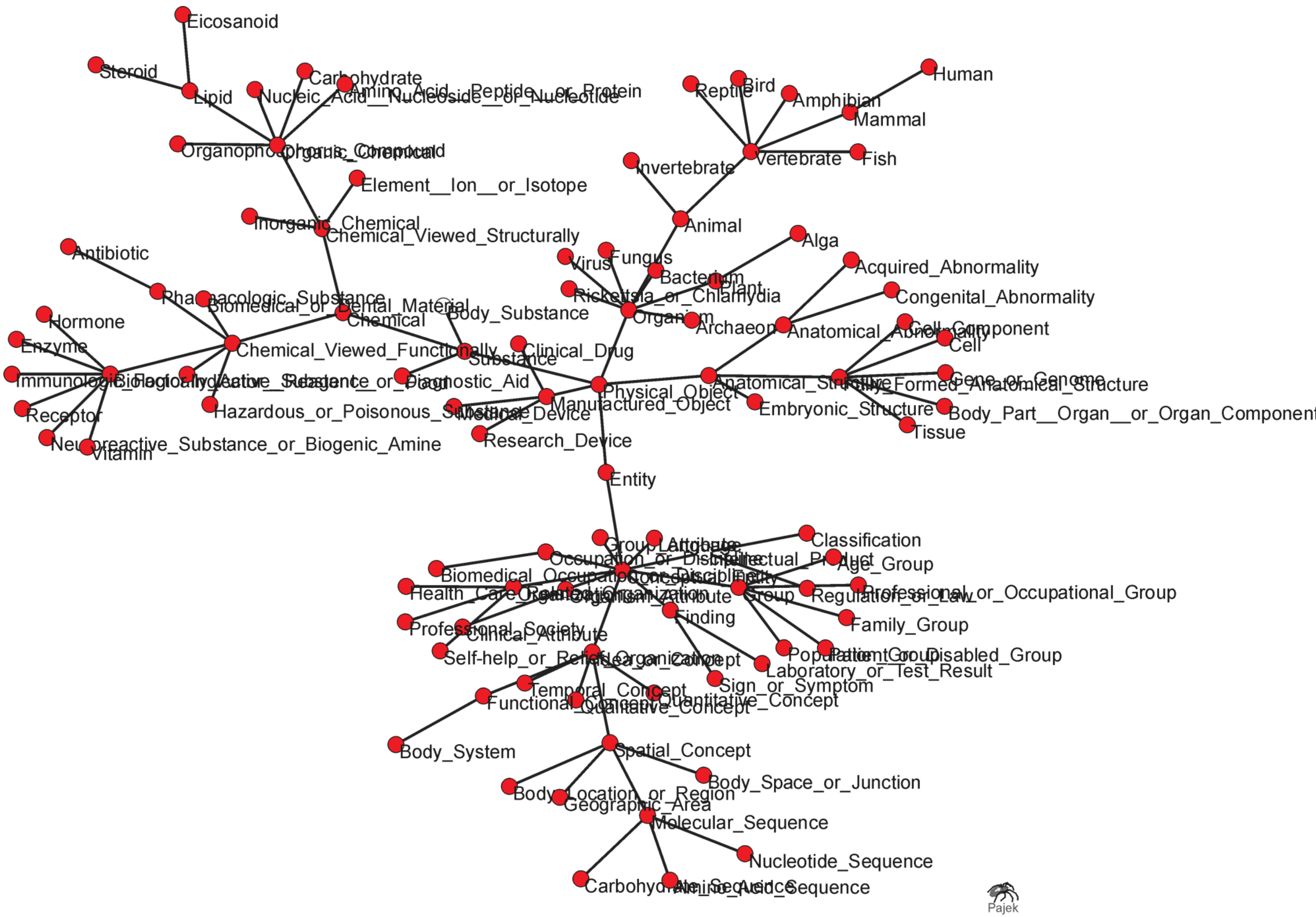
- Key question – assignment of concepts to modules
- Module – information about a subtopic – can stand for itself – concepts within module are semantically connected
- The resulting module – weighted graph G=(C,D,w)
- Dependencies
  - Reflected in definitions of O
  - Implied by the intuitive understanding of concepts and a background knowledge about domain
  - Different structures
    - Subclass relations between classes
    - Other relations (range, domain restrictions ...)

Partitioning method

- Decomposition of larger ontologies to smaller modules
- Consists of three steps
  1. Create ontology graph known as weighted or dependency – two tasks
     - Extraction ontology source file
     - Determine strength of relations
  2. Identification of modules
     - Determine concept Island
  3. Optimization of partition
     - Assign isolated concepts

Create dependency graph

- Create semantic network in which concepts are represented by nodes
- relations between concepts
- On the following figure – class hierarchy graph of the part of UMLS semantic network


Eicosanoid
Steroid
Lipid
Carbohydrate
Nucleic_Acid__Nucleoside__or_Nucleotide
Amino_Acid__Peptide__or_Protein
Organic_Chemical
Element__Ion__or_Isotope
Inorganic_Chemical
Chemical_Viewed_Structurally
Antibiotic
Pharmacologic_Substance
Biomedical_or_Dental_Material
Chemical
Body_Substance
Hormone
Enzyme
Chemical_Viewed_Functionally
Substance
Clinical_Drug
Receptor
Hazardous_or_Poisonous_Substance
Diagnostic_Aid
Medical_Device
Research_Device
Vitamin
Neuroreactive_Substance_or_Biogenic_Amine
Manufactured_Object
Physical_Object
Entity
Reptile
Bird
Human
Amphibian
Mammal
Vertebrate
Fish
Invertebrate
Animal
Alga
Fungus
Virus
Bacterium
Plant
Rickettsia_or_Chlamydia
Organism
Archaeon
Acquired_Abnormality
Congenital_Abnormality
Anatomical_Abnormality
Cell_Component
Cell
Anatomical_Structure
Gene_or_Genome
Fully_Formed_Anatomical_Structure
Embryonic_Structure
Body_Part__Organ__or_Organ_Component
Tissue
Classification
Language
Occupation_or_Discipline
Intellectual_Product
Age_Group
Biomedical_Occupation_or_Discipline
Conceptual_Entity
Health_Care_Related_Organization
Organism_Attribute
Group
Regulation_or_Law
Professional_or_Occupational_Group
Finding
Family_Group
Professional_Society
Clinical_Attribute
Population_Group
Patient_or_Disabled_Group
Self-help_or_Relief_Organization
Idea_or_Concept
Laboratory_or_Test_Result
Sign_or_Symptom
Temporal_Concept
Functional_Concept
Qualitative_Concept
Quantitative_Concept
Body_System
Spatial_Concept
Body_Space_or_Junction
Body_Location_or_Region
Geographic_Area
Molecular_Sequence
Nucleotide_Sequence
Carbohydrate_Sequence
Amino_Acid_Sequence
Pajek

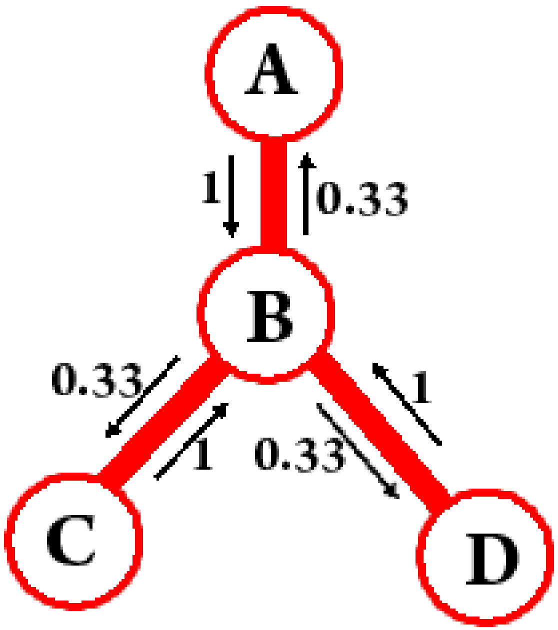
UMLS - Unified Medical Language System

- developed by the US National Library of Medicine (1986)
- integrates over 2 million names for some 900 000 concepts from more than 60 families of biomedical vocabularies
- Three parts
  1. Metathesaurus
     - Organized by meaning, it doesn´t create ontology itself
  2. Semantic network
     - Provides semantic relationships among concepts
  3. Special Lexicon
     - contains syntactic, morphological and orthographic dictionary

Determine strength of relations

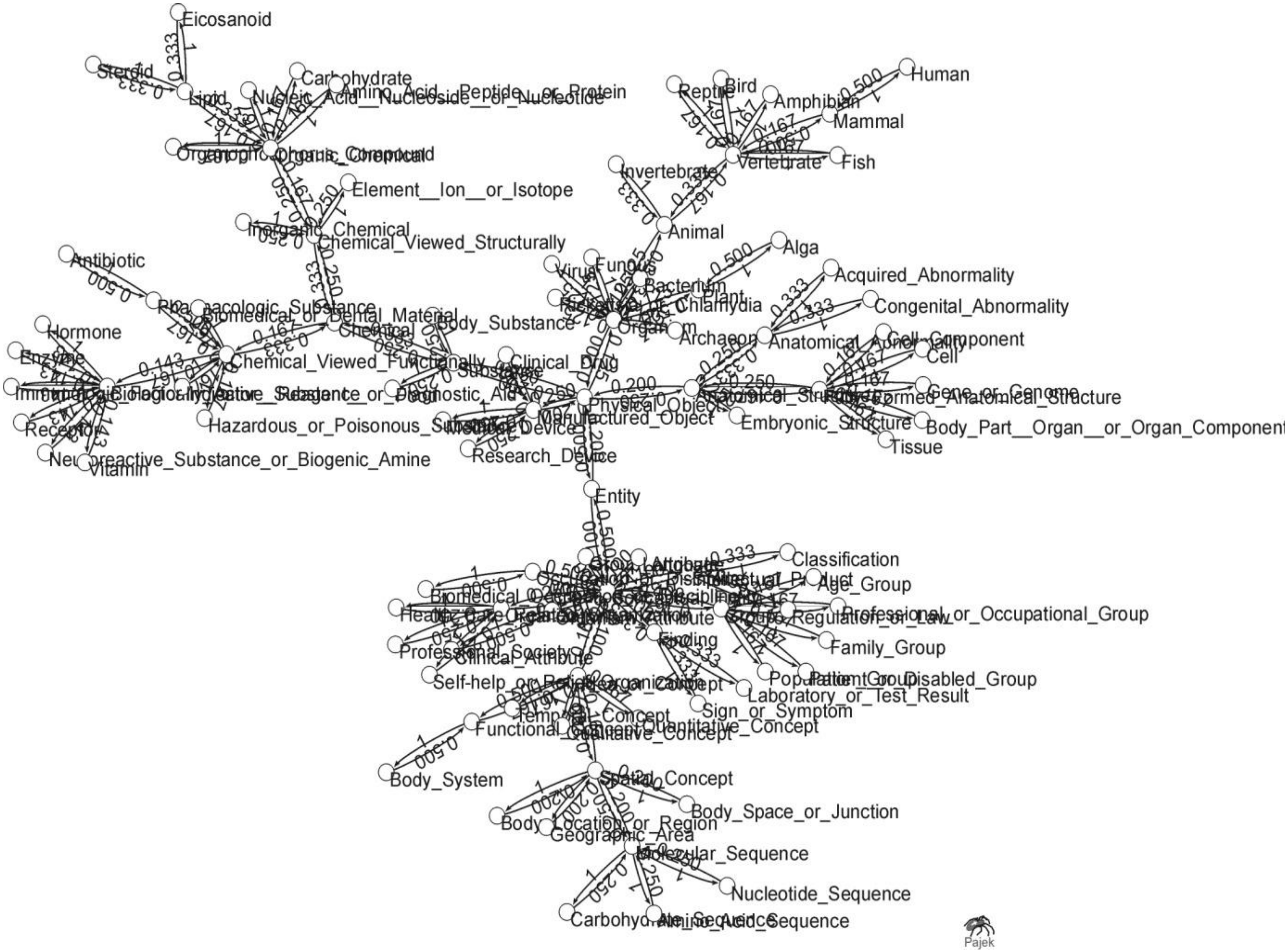
- The structure of dependency graph is used to determine strength among concepts (nodes)
- Using social network theory by computing the proportional strength
- p_{ij} of a connection between a node c_i and c_j – importance of a link from one node to other based on the number of connections a node has

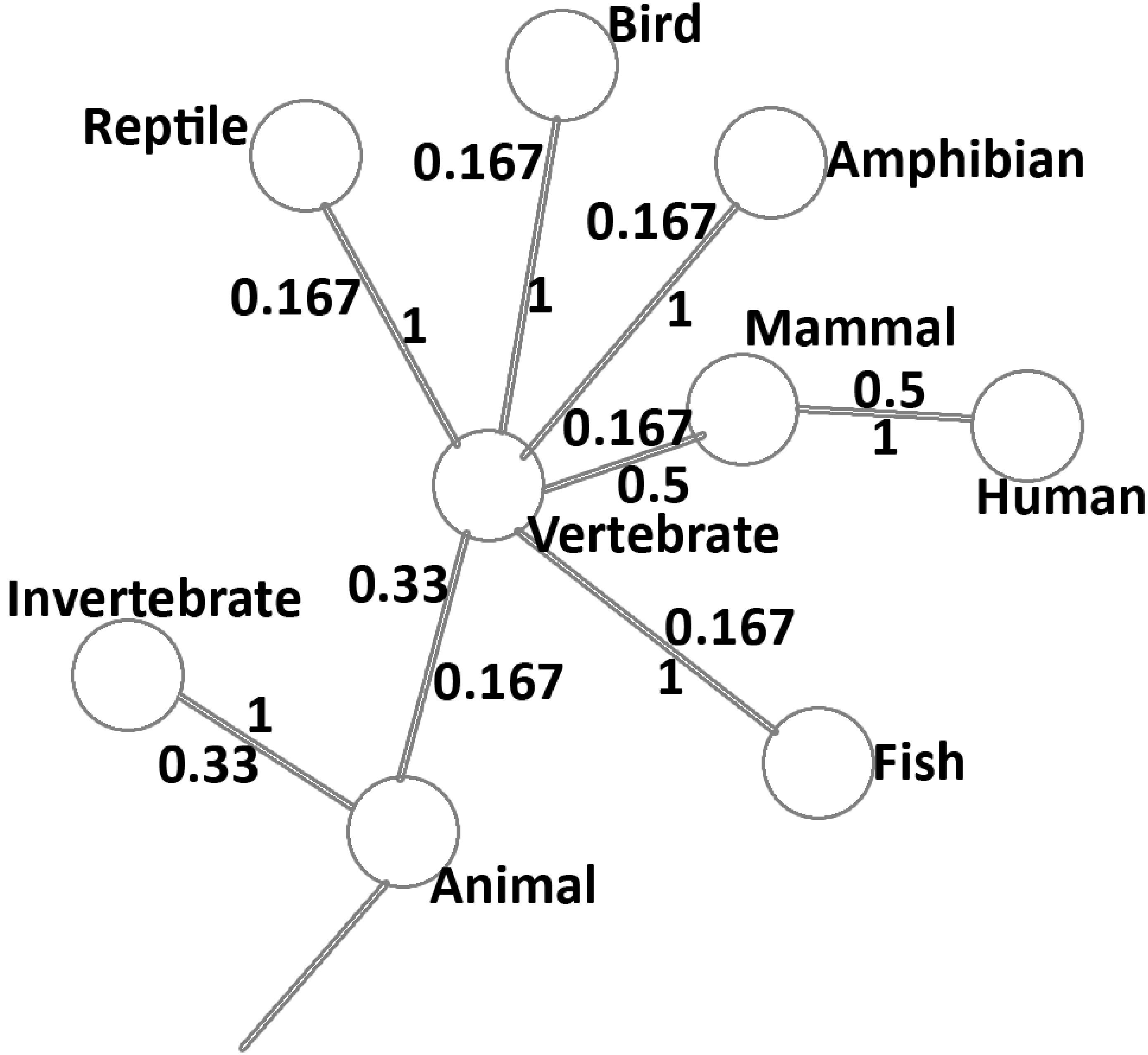
General example of a proportional strength



- Four nodes A, B, C, D
- A → B , ps = 1
  - A has one connection (B)
- B → A , ps = 0.33
  - B has three connections (A,C,D)

Therefore an assymetric connection among concepts

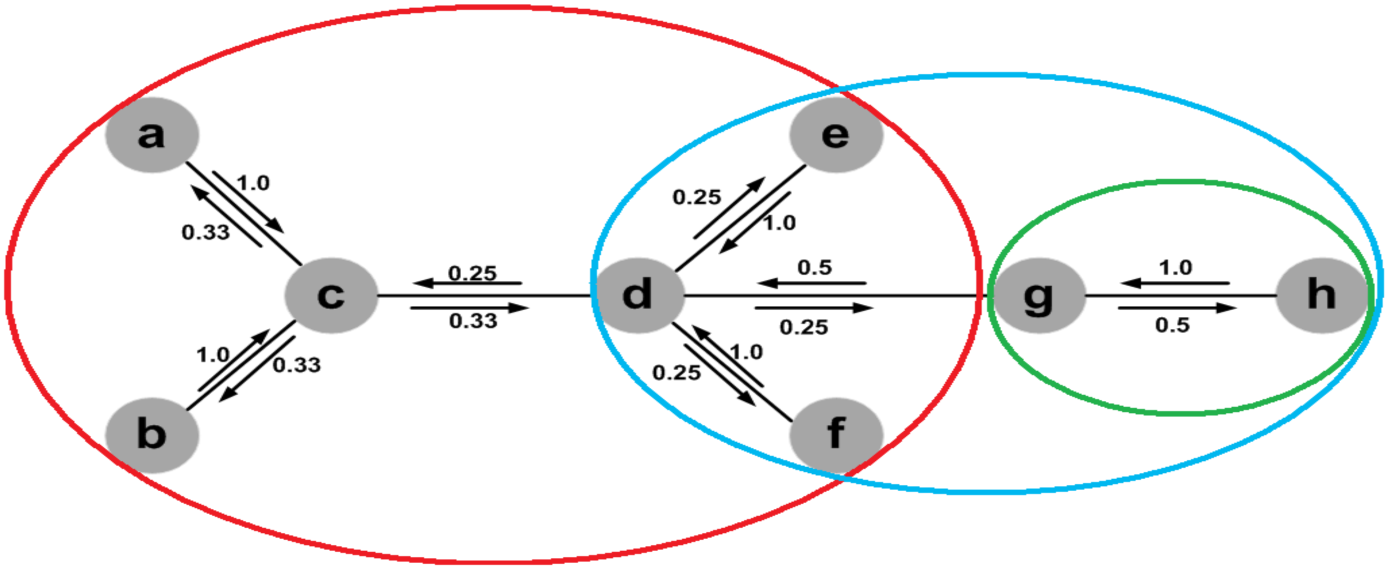

Eicosanoid
Steroid
Lipid
Carbohydrate
Amino_Acid__Peptide__or_Protein
Nucleic_Acid__Nucleoside__or_Nucleotide
Element__Ion__or_Isotope
Inorganic_Chemical
Chemical_Viewed_Structurally
Antibiotic
Pharmacologic_Substance
Biomedical_or_Dental_Material
Body_Substance
Chemical
Hormone
Enzyme
Chemical_Viewed_Functionally
Clinical_Drug
Hazardous_or_Poisonous_Substance
Receptor
Vitamin
Research_Device
Physical_Object
Entity
Human
Reptile
Bird
Amphibian
Mammal
Vertebrate
Fish
Invertebrate
Animal
Alga
Fungus
Virus
Bacterium
Plant
Archaeon
Acquired_Abnormality
Congenital_Abnormality
Cell_Component
Cell
Gene_or_Genome
Embryonic_Structure
Body_Part__Organ__or_Organ_Component
Tissue
Classification
Age_Group
Professional_or_Occupational_Group
Family_Group
Laboratory_or_Test_Result
Sign_or_Symptom
Finding
Professional_Society
Body_System
Body_Space_or_Junction
Body_Location_or_Region
Geographic_Area
Molecular_Sequence
Nucleotide_Sequence
Pajek


Bird
Reptile
0.167
Amphibian
0.167
0.167
1
1
1
Mammal
0.5
0.167
1
0.5
Human
Vertebrate
0.33
Invertebrate
0.167
1
0.167
1
0.33
Fish
Animal

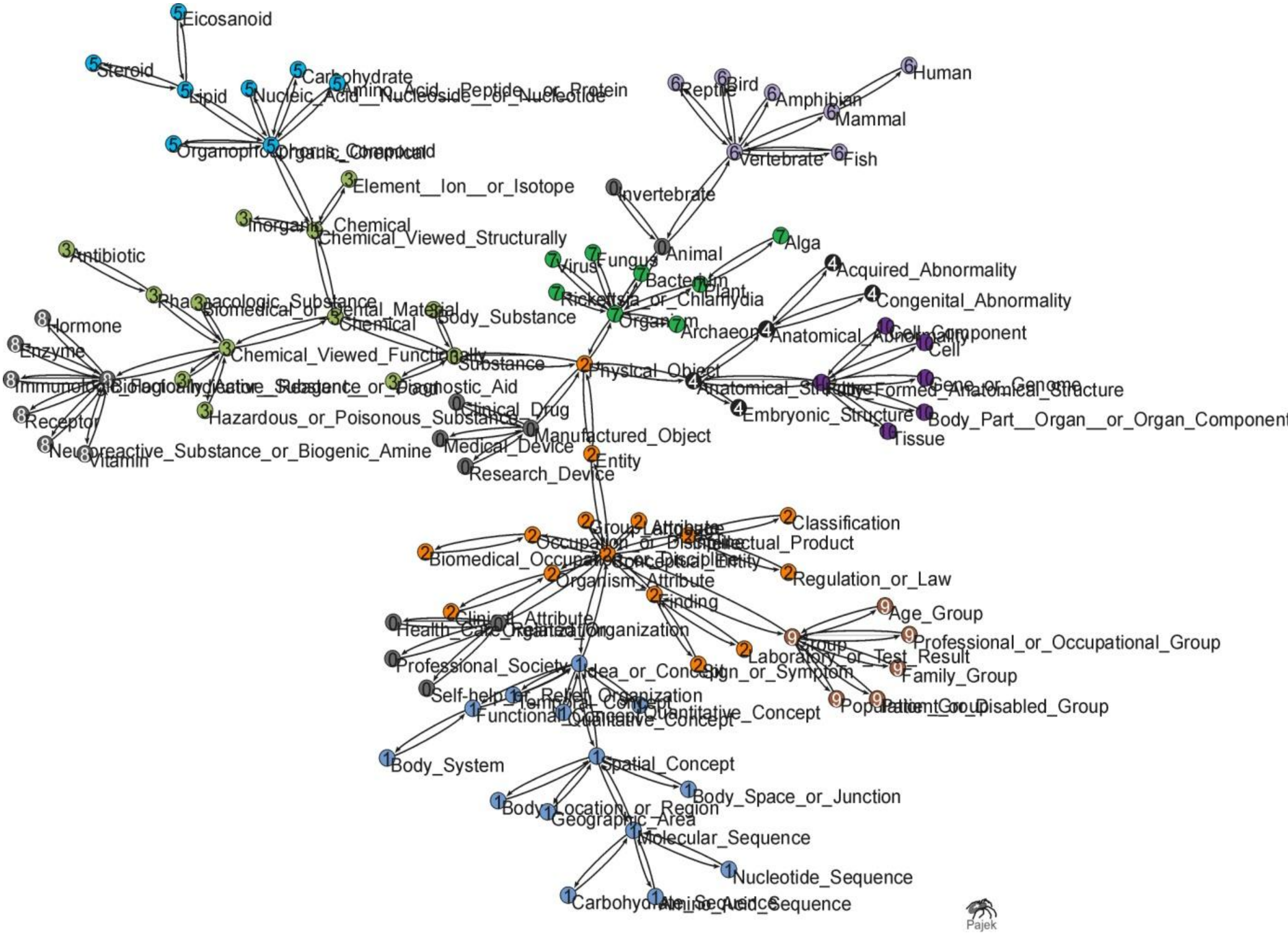
Identification of modules

- Using the algorithm to compute all maximal Line Islands
- One Island represents One Module
- A set of vertices I ⊆ C is a Line Island in dependency graph ⊆ G=(C,D,w) if and only if existing connected subgraph and lines inside the subgraph are more strongly related among them than with neighboring vertices – **Maximal Spanning tree T** – his weight is bigger than the weight of every other spanning tree
- It is necessary to determine the upper and lower bound – size of module

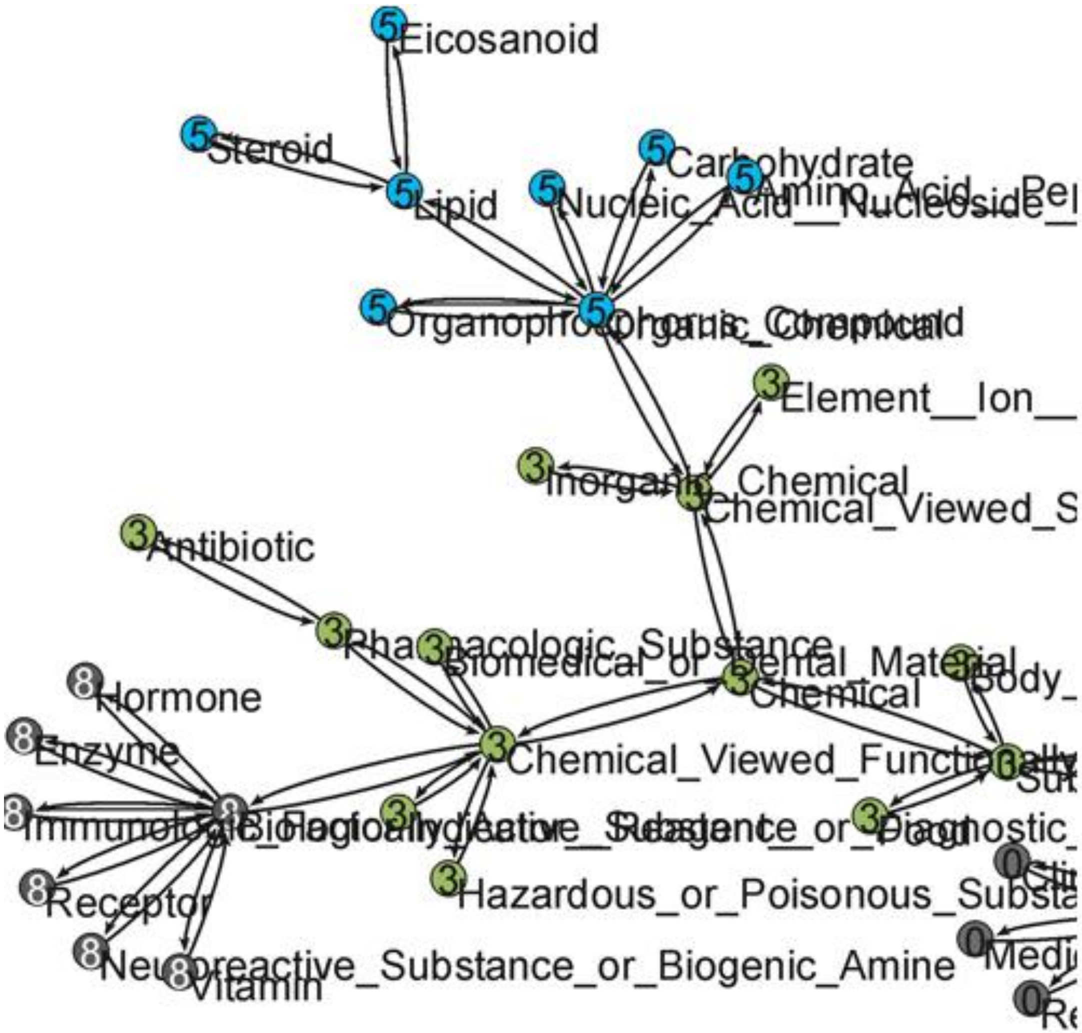
General description of Line Island



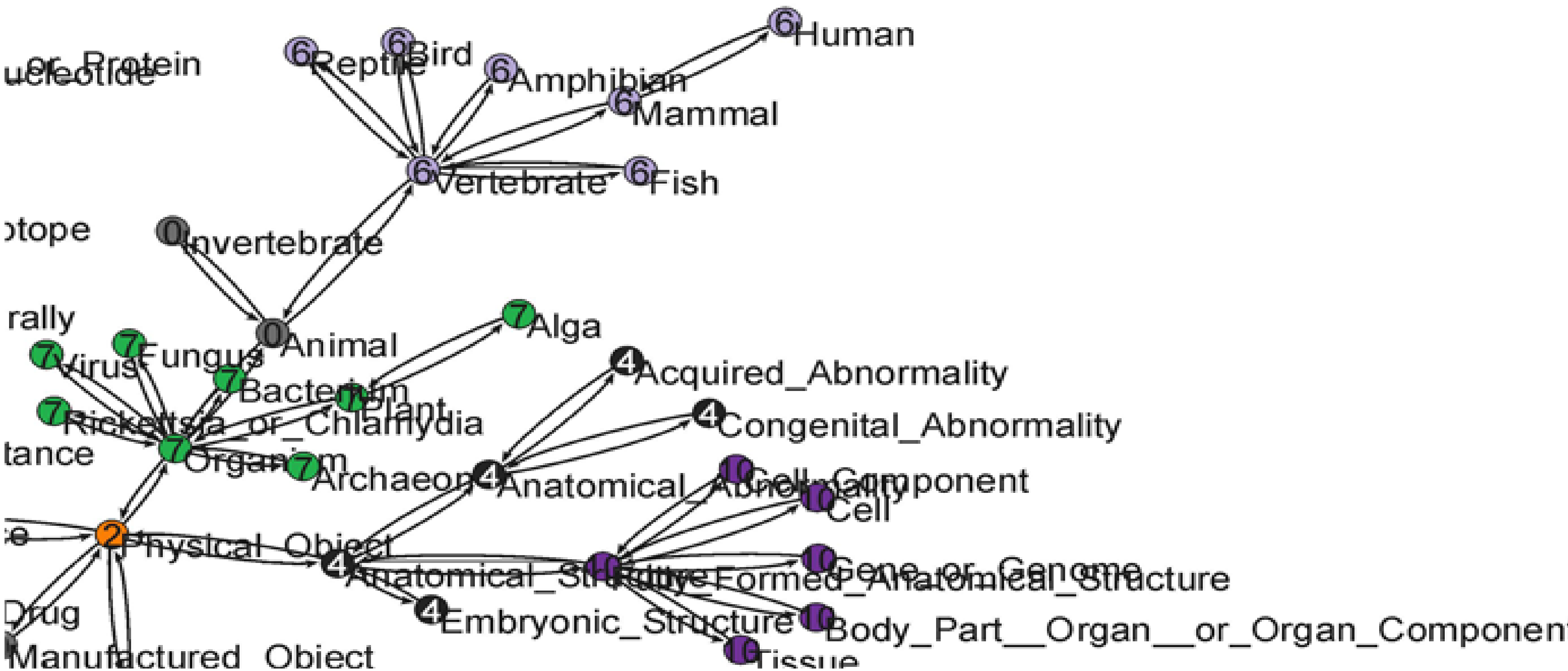
- Napr:
  - {a,b,c,d,e,f} – is not LI („Line Island") - PS between **c** a **d** is 0.33 but between **g** a **d** 0.5, PS is bigger
  - {g,h} – is LI – maximal value of an input and output connection is 0.5 but this isn't the maximal spanning tree
  - {d,e,f,g,h} - LI with the maximal spanning tree



Eicosanoid
Steroid
Lipid
Carbohydrate
Amino_Acid__Peptide__or_Protein
Organic_Chemical
Element__Ion__or_Isotope
Inorganic_Chemical
Chemical_Viewed_Structurally
Antibiotic
Pharmacologic_Substance
Chemical
Body_Substance
Hormone
Enzyme
Chemical_Viewed_Functionally
Substance
Receptor
Hazardous_or_Poisonous_Substance
Vitamin
Clinical_Drug
Manufactured_Object
Medical_Device
Research_Device
Entity
Physical_Object
Human
Bird
Amphibian
Mammal
Vertebrate
Fish
Invertebrate
Animal
Alga
Fungus
Bacterium
Rickettsia_or_Chlamydia
Archaeon
Acquired_Abnormality
Congenital_Abnormality
Cell_Component
Cell
Embryonic_Structure
Body_Part__Organ__or_Organ_Component
Tissue
Classification
Intellectual_Product
Conceptual_Entity
Regulation_or_Law
Organism_Attribute
Finding
Age_Group
Clinical_Attribute
Group
Professional_or_Occupational_Group
Laboratory_or_Test_Result
Family_Group
Professional_Society
Sign_or_Symptom
Self-help_or_Relief_Organization
Quantitative_Concept
Qualitative_Concept
Patient_or_Disabled_Group
Body_System
Spatial_Concept
Body_Space_or_Junction
Body_Location_or_Region
Geographic_Area
Molecular_Sequence
Nucleotide_Sequence
Pajek

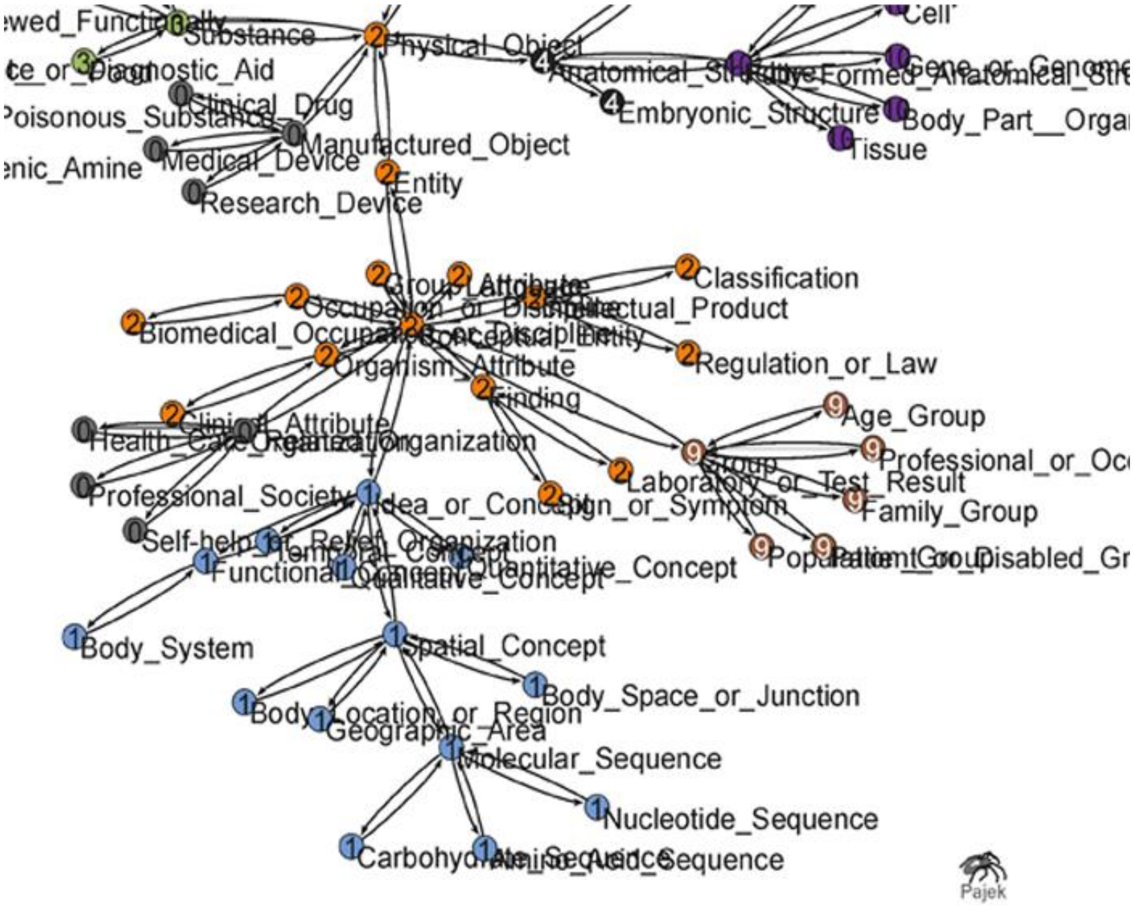


- 3. Chemical
- 5. Organic chemical
- 8. Biologically Active Substance



- 4. Anatomical Structure
- 6. Vertebrate
- 7. Organism
- 10. Fully Formed Anatomical Structure

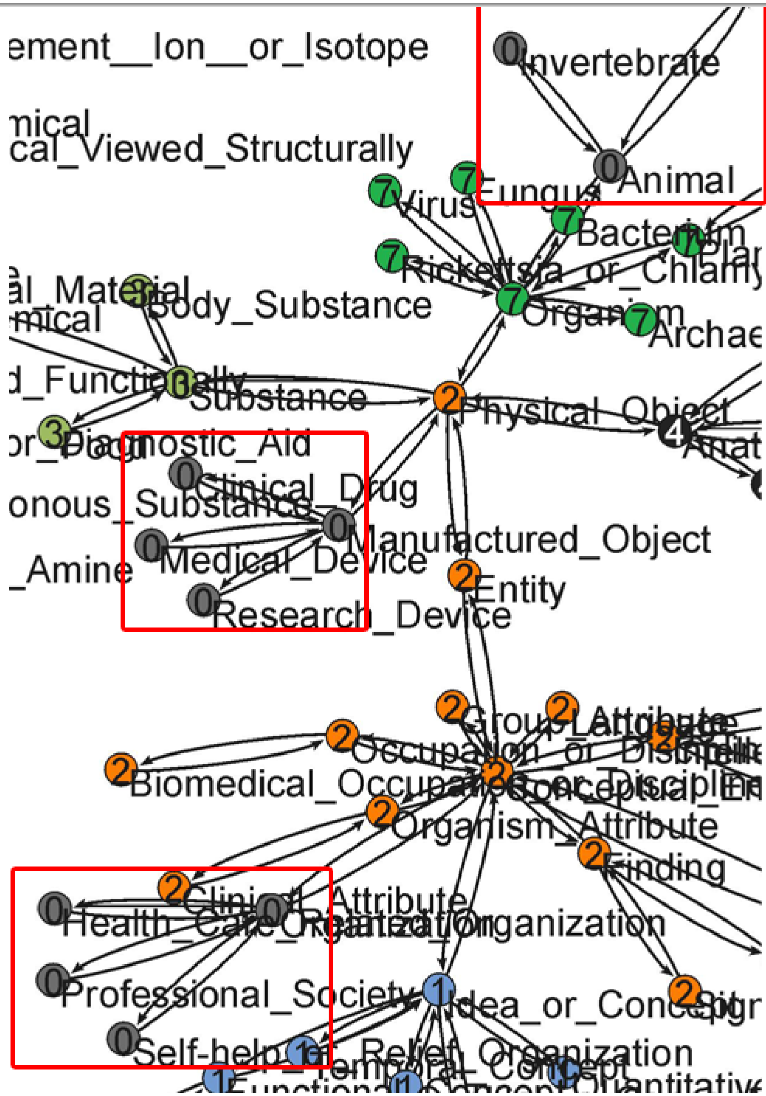
- 1. Idea or Concept
- 2. Entity
- 9. Group



Isolated concepts

- Islands **$\alpha(c)=i$**
- If **$\alpha(c)=o$** - concept can be assigned to any module
- this situation may happen when nodes cannot be assigned to islands – these concepts are known as isolated (unassigned) concepts

10 isolated concepts

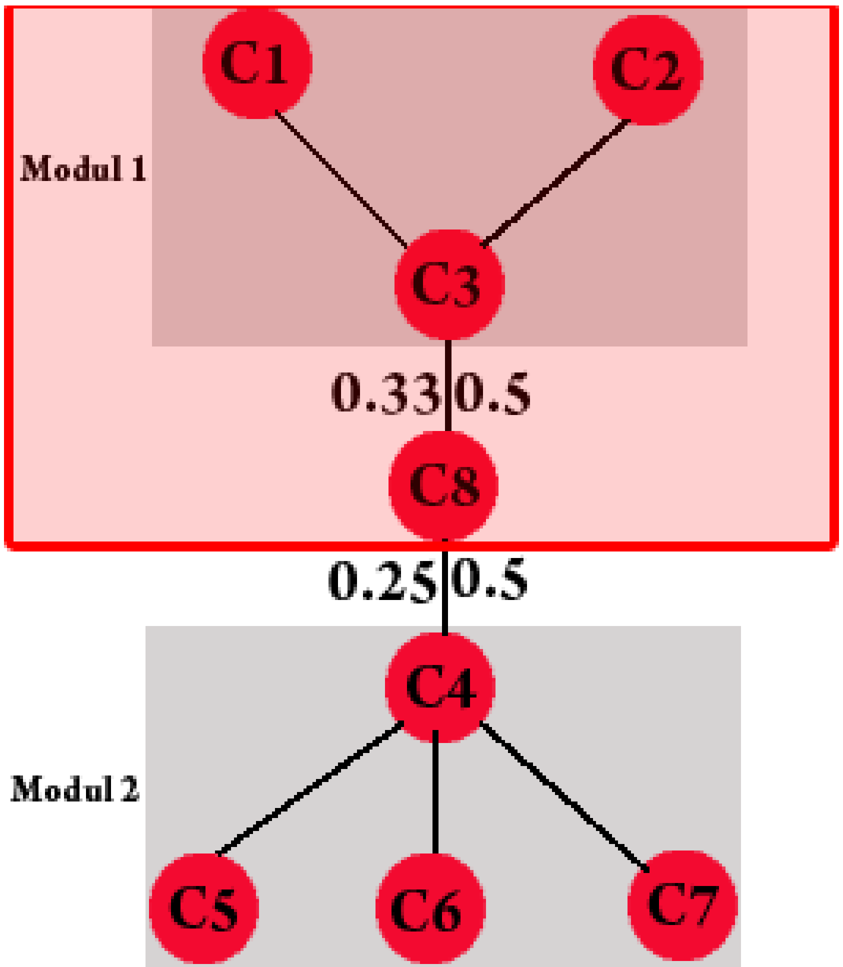


- 4 nodes related with Organizations
- 4 nodes related with Manufactured Object
- 2 nodes related with Animal and Invertebrate

Optimization of partition – assign isolated concepts

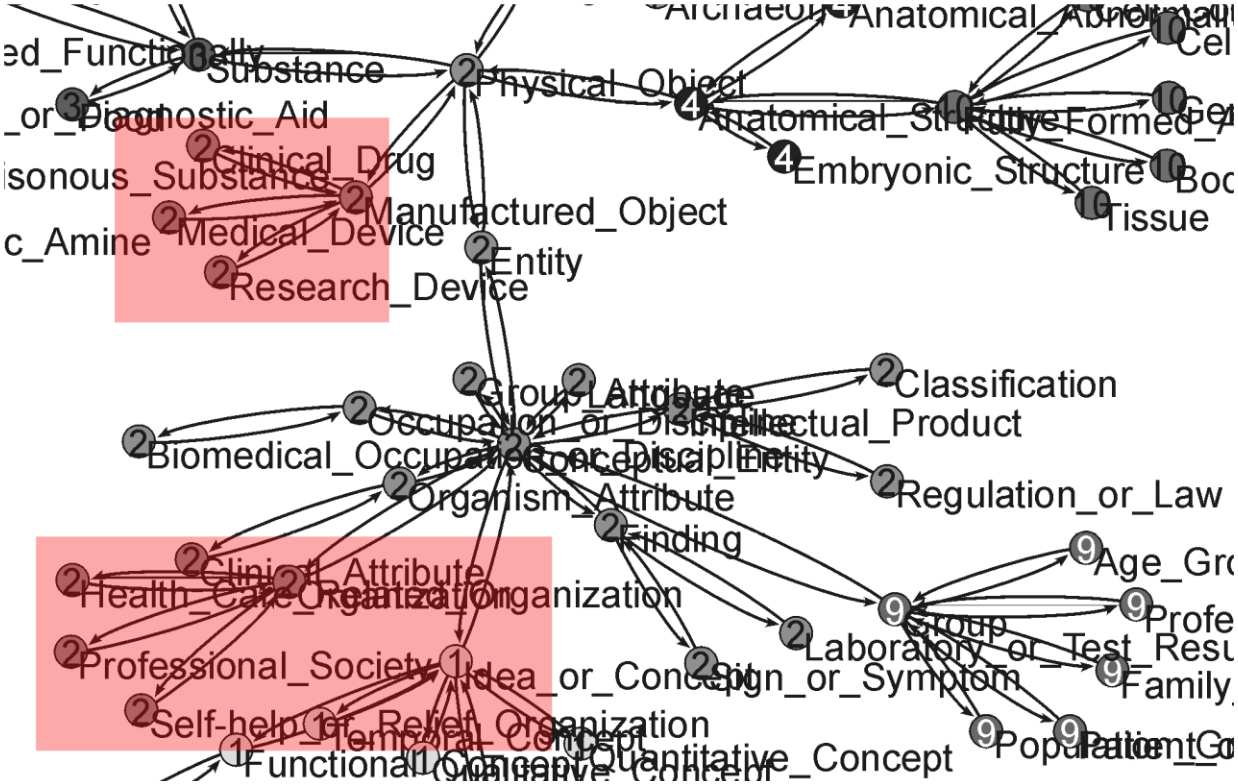
- Leftover nodes can occur in different places in the graph
- Isolated nodes are assigned to other nodes – the assignement is based on the strength of relations to nodes, that are already assigned to an existing module – the nodes are assigned to the Island of a neighboring node which has the strongest relations among all neighboring nodes around the isolated nodes

General example of assignment isolated nodes



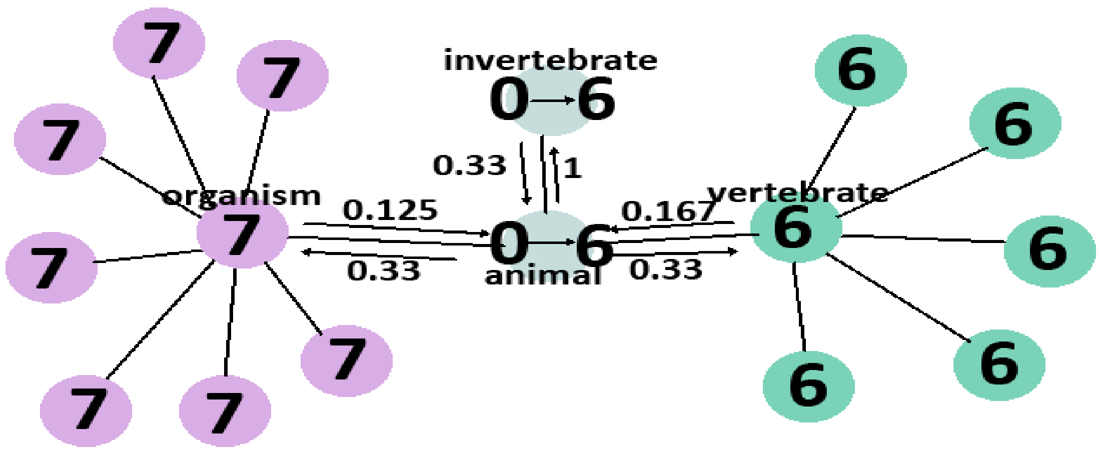
- $C_3 \longrightarrow C_8$ … 0.33 0.5 = 0.85
- $C_4 \longrightarrow C_8$ … 0.25 0.5 = 0.75

The assignment of 10 isolated concepts



- 4 nodes related with Organizations are assigned to module 2
- 4 nodes related with Manufactured Object are assigned to module 2 too

The assignment of 10 isolated concepts



- 2 nodes related with Animal and Invertebrate are assigned to module 6
- calculation:
  - 0.167+0.33 > 0.125+0.33

Evaluation of an algorithm

- Main problem - necessary to determine the size of modules (upper and lower limit) – bad choice of the bound leads to high number of unassigned nodes – after the assignement leftover nodes – quite large modules with little internal coherence
- Iterative algorithm can eliminate this issue

Iterative algorithm

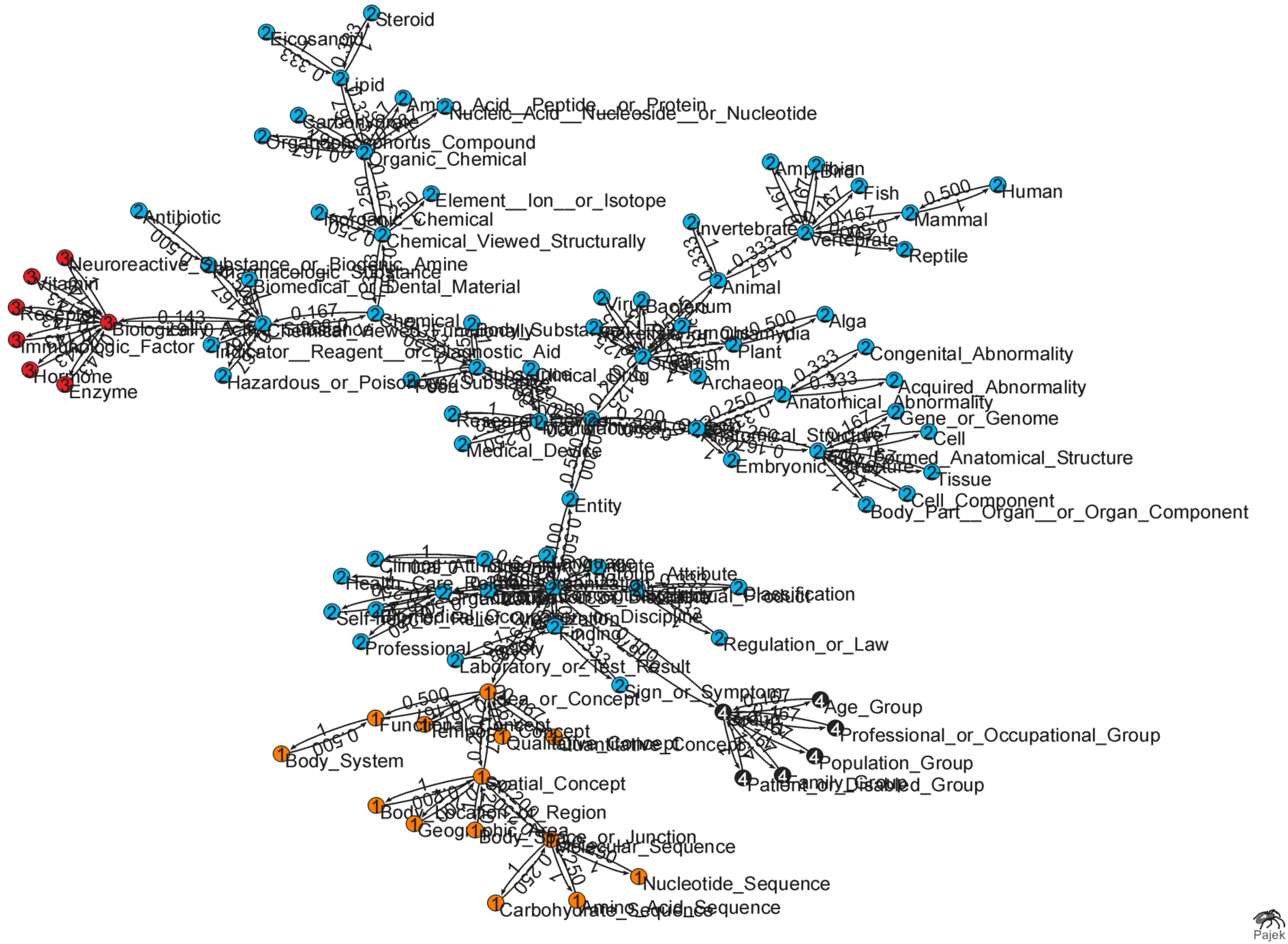
- Idea – is not to prescribe the size of modules
- To set the lower bound to **1** and the upper to **s-1**, s – size of the complete ontology
- Choosing a limit that is just one below the size of complete ontology does not further restrict the selection of islands – this way – the most natural grouping of concepts
- But – it can happen that nodes cannot be assigned to Islands
- the result – Islands differ in size, often large modules that cover most of the ontology – therefore - iteratively apply the algorithm

Example of Iterative algorithm

- The first step – to determine the upper limit of 20 for the size of modules
- The module is relatively small therefore the algorithm only needs three iterations

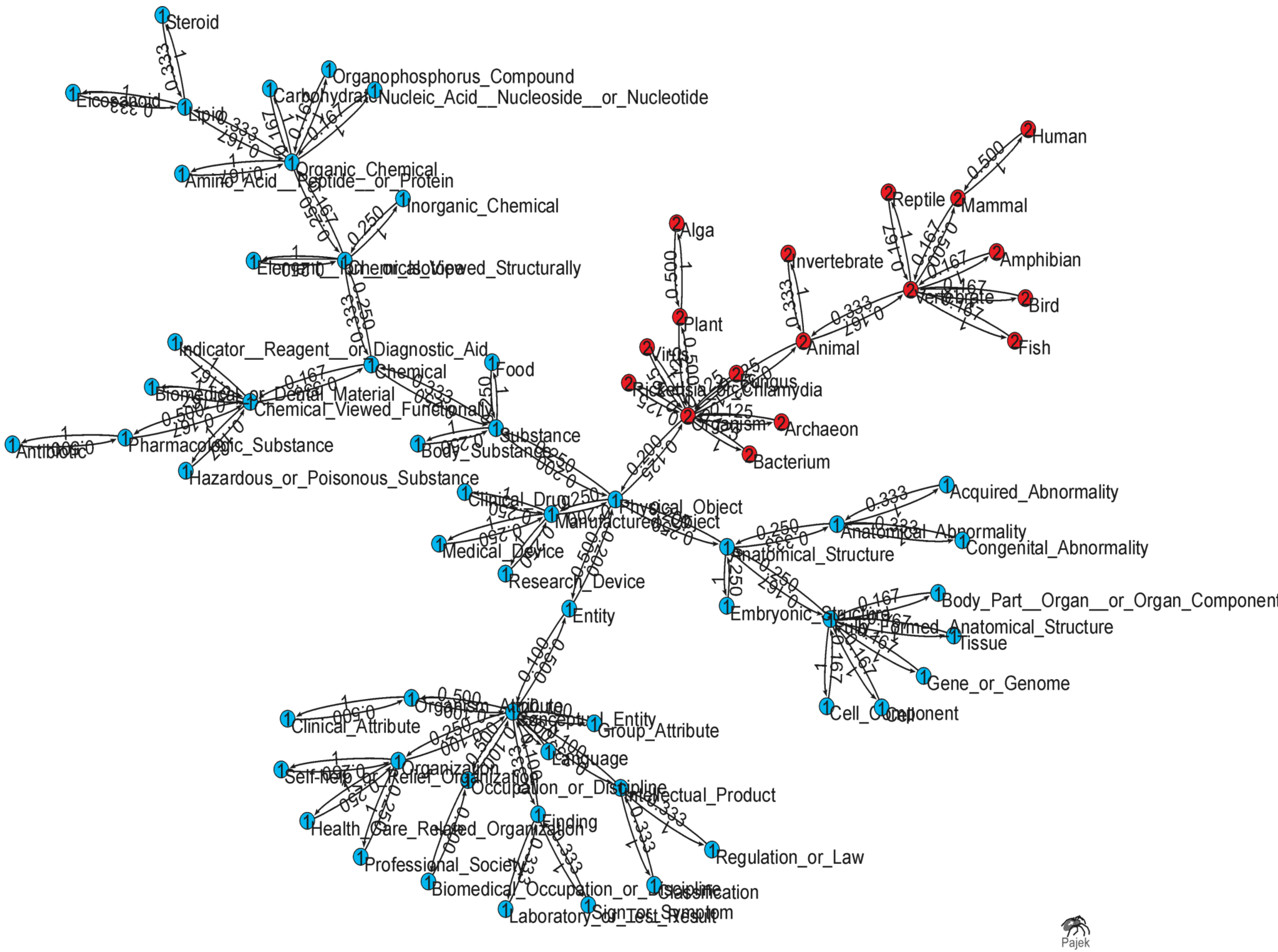
First iteraion

- The algorithm generates only four modules
- Three of them are smaller than determined limit (20)
- Modules
  - Biological active substance
  - Idea or Concept
  - Different Age Group
  - Leftover part of ontology represents large module
- Module Biological active substance could be included in a larger module
- The other two contain concepts that are related and sufficiently different from other concepts


Steroid
Eicosanoid
Lipid
Amino_Acid__Peptide__or_Protein
Nucleic_Acid__Nucleoside__or_Nucleotide
Carbohydrate
Organophosphorus_Compound
Organic_Chemical
Element__Ion__or_Isotope
Inorganic_Chemical
Chemical_Viewed_Structurally
Antibiotic
Neuroreactive_Substance_or_Biogenic_Amine
Pharmacologic_Substance
Vitamin
Biomedical_or_Dental_Material
Receptor
Immunologic_Factor
Hormone
Enzyme
Chemical
Indicator__Reagent__or_Diagnostic_Aid
Hazardous_or_Poisonous_Substance
Medical_Device
Amphibian
Bird
Fish
Human
Mammal
Invertebrate
Vertebrate
Reptile
Animal
Bacterium
Alga
Plant
Organism
Archaeon
Congenital_Abnormality
Acquired_Abnormality
Anatomical_Abnormality
Gene_or_Genome
Cell
Fully_Formed_Anatomical_Structure
Tissue
Embryonic_Structure
Cell_Component
Body_Part__Organ__or_Organ_Component
Entity
Classification
Regulation_or_Law
Professional_Society
Laboratory_or_Test_Result
Finding
Sign_or_Symptom
Idea_or_Concept
Functional_Concept
Qualitative_Concept
Quantitative_Concept
Age_Group
Professional_or_Occupational_Group
Population_Group
Patient_or_Disabled_Group
Body_System
Spatial_Concept
Body_Location_or_Region
Geographic_Area
Molecular_Sequence
Nucleotide_Sequence
Carbohydrate_Sequence
Amino_Acid_Sequence
Pajek

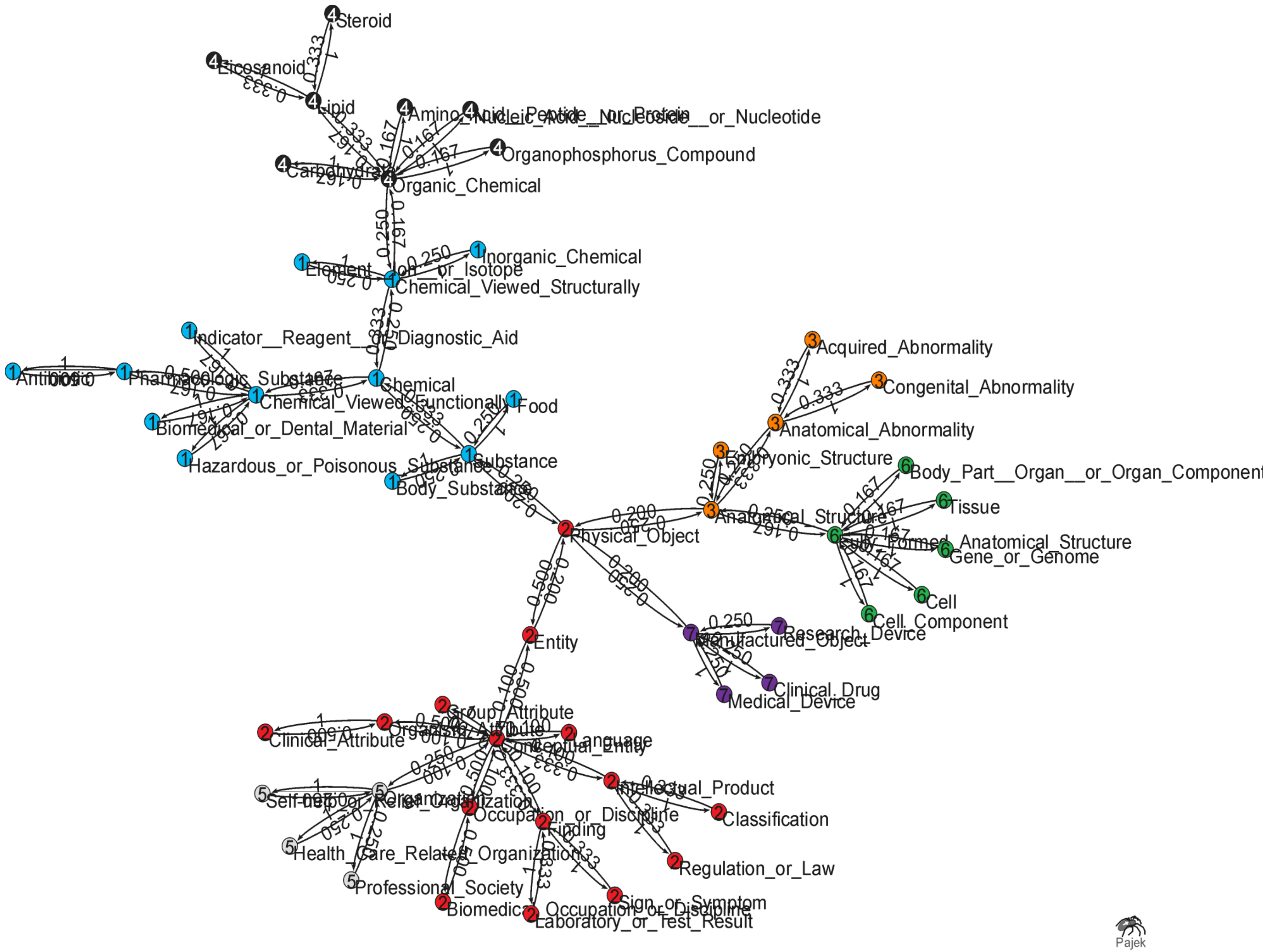
Second iteration

- After removing the modules found in the first step, the algorithm generates new module Organism


Steroid
Eicosanoid
Lipid
Organophosphorus_Compound
Nucleic_Acid__Nucleoside__or_Nucleotide
Organic_Chemical
Amino_Acid__Peptide__or_Protein
Inorganic_Chemical
Chemical_Viewed_Structurally
Indicator__Reagent__or_Diagnostic_Aid
Chemical
Food
Biomedical_or_Dental_Material
Chemical_Viewed_Functionally
Antibiotic
Pharmacologic_Substance
Substance
Body_Substance
Hazardous_or_Poisonous_Substance
Clinical_Drug
Physical_Object
Manufactured_Object
Medical_Device
Research_Device
Entity
Human
Reptile
Mammal
Alga
Invertebrate
Amphibian
Vertebrate
Bird
Plant
Animal
Fish
Virus
Fungus
Rickettsia_or_Chlamydia
Organism
Archaeon
Bacterium
Acquired_Abnormality
Anatomical_Abnormality
Congenital_Abnormality
Anatomical_Structure
Embryonic_Structure
Body_Part__Organ__or_Organ_Component
Fully_Formed_Anatomical_Structure
Tissue
Gene_or_Genome
Clinical_Attribute
Group_Attribute
Language
Organization
Occupation_or_Discipline
Intellectual_Product
Health_Care_Related_Organization
Finding
Regulation_or_Law
Professional_Society
Sign_or_Symptom
Pajek

Third iteration

- produces a partition of the remaining concepts into Islands – all of the required size – ending the iteration
- Result – seven modules
  - Entity
  - Organization
  - Device
  - Anatomical Structure
  - Fully Formed Anatomical Structure
  - Substance
  - Organic Chemical


Steroid
Eicosanoid
Lipid
Nucleic_Acid__Nucleoside__or_Nucleotide
Organophosphorus_Compound
Organic_Chemical
Inorganic_Chemical
Ion_or_Isotope
Chemical_Viewed_Structurally
Indicator__Reagent__or_Diagnostic_Aid
Chemical
Chemical_Viewed_Functionally
Food
Biomedical_or_Dental_Material
Hazardous_or_Poisonous_Substance
Substance
Body_Substance
Acquired_Abnormality
Congenital_Abnormality
Anatomical_Abnormality
Embryonic_Structure
Body_Part__Organ__or_Organ_Component
Tissue
Gene_or_Genome
Cell
Cell_Component
Physical_Object
Entity
Research_Device
Clinical_Drug
Medical_Device
Group_Attribute
Clinical_Attribute
Intellectual_Product
Classification
Regulation_or_Law
Health_Care_Related_Organization
Professional_Society
Occupation_or_Discipline
Finding
Sign_or_Symptom
Laboratory_or_Test_Result
Pajek

Result of iteration algorithm

- Most of these modules make sense
- Only two modules are arguable
  - Separation of Fully Formed Anatomical Structures from Anatomical Structures
  - Separation of Organic Chemical from Substance

Conclusion

- Algorithm generates modules, that fulfill our expectations to a certain extent
- Sometimes subtrees, that could be considered to form one module are further split, even if subtree does not exceed the upper size limit
- In spite of the fact that iterative algorithm doesn’t require determination of the upper limit of module, generated modules may become too small to reflect the real world

The future work

- The aim
  - Elimination of possible mistakes
  - Optimization of algorithm of decomposition method

References

- Mathieu d’Aquin, Anne Schlicht, HeinerStuckenschmidt, and Marta Sabou, "Ontology Modularization for Knowledge Selection: Experiments and Evaluations".
- Heiner Stuckenschmidth, Michel Klein, “Towards structure-Based Partitioning o Large Ontologies“ Vrije Universiteit Amsterdam
- Heiner Stuckenschmidth, Anne Schlicht, “Structure-Based Partitioning o Large Ontologies“ Universitat Mannheim, Germany
- Stefano Spaccapietra, "Report on Modularization of Ontologies," Institute of Computer Science, Austria
- http://www.ncbi.nlm.nih.gov/pubmed/9082131
- Peter Lesný, Ján Vejvalka, “UMLS pro Medigrid”
- http://nar.oxfordjournals.org/cgi/content/abstract/32/suppl_1/D267
- Paul Doran, “Ontology reuse via ontology modularisation“ Department of Computer Science, University of Liverpool
- Heiner Stuckenschmidt and Michel Klein, “Integrity and Change in Modular Ontologies“ Vrije Universiteit Amsterdam
- Bernardo Cuenca Grau, Bijan Parsia, Evren Sirin, Aditya Kalyanpur, “Modularizing OWL Ontologies“ University of Maryland at College Park

Thank for your attention