MŰEGYETEM 1782

Budapest University of Technology and Economics
Faculty of Electrical Engineering and Informatics
Department of Telecommunications and Informatics

Evolution of the Hierarchical Network of Medical Subject Headings

Scientific Students' Association Report

Author:

Daniel Zagyva

Advisors:

Gergely Palla, MTA-ELTE Statistical and Biological Physics Research Group

Gergő Barta, PhD Student

Table of Contents

Kivonat

Sok természetben és társadalomban megtalálható komplex rendszerre jellemző a hierarchikus szerveződés, számos kutatás foglalkozott már különböző technológiai, ökológiai, biológiai, társadalmi hálózatok hierarchikus tulajdonságával. Több hálózatelméleti modell is a valós rendszerekben megfigyelt növekedési tulajdonságra épít, és egy érdekes, eddig kevésbé vizsgált kutatási kérdés a hierarchikus hálózatok időfejlődésének leírása.

Ebben a munkában különböző adatelemzési módszerekkel vizsgáljuk az NCBI MeSH tárgyszavak hierarchikus hálózatának időfejlődését, mely 16 különböző, évenként frissülő hierarchiából áll kategóriától függően: pl.: Anatómia, Betegségek, Vegyszerek, Gyógyszerek stb.

Mindegyik hierarchia megfeleltethető egy-egy irányított körmentes gráfnak (DAG - directed acyclic graph), melynek irányított élei a hierarchia felsőbb szintjei felől mutatnak az alsóbb szintek felé. Mivel újabb és újabb tárgyszavak jellennek meg az adatbázisban, a MeSH hierarchiák növekednek az idő során.

A kutatási eredményeink szerint nem csak a növekedés, hanem a meglévő kapcsolatok közötti átrendeződés is nagy szerepet játszik a hierarchiák formálásában. A dolgozatban ennek az időfejlődésnek számos statisztikai jellemzőjét vizsgáljuk, valamint a MeSH hierarchiák általános tulajdonságait írjuk le.

Habár a kutatási eredményeink a MeSH tárgyszavak hálózatára korlátozódik, valószínű, hogy a leírt jellemzők egy része általánosítható, és megjelenhet a legtöbb hierarchikus hálózat időfejlődése során.

Abstract

Hierarchical organization is a prevalent feature of many complex networks appearing in nature and society. Recent research works on hierarchy include technological networks, ecological systems, social interactions and even neural networks.

Several proposed models exist in network theory which are based on the observed growth patterns of real networks. A related interesting, yet less studied question is how does a hierarchical network evolve in time?

Here we take a data-driven approach and examine the time evolution of the network between the Medical Subject Headings (MeSH) provided by the NCBI, which are organized into 16 different, yearly updated hierarchies such as Anatomy, Diseases, Chemicals and Drugs, etc. The natural representation of these hierarchies is given by directed acyclic graphs (DAGs), composed of links pointing from nodes higher in the hierarchy towards nodes in lower levels. Due to the appearance of new MeSH terms, the MeSH hierarchies are growing in time.

According to our results, not only growing but also restructuring of the already existing connections plays an important role as well in forming the shape of the DAGs. We examine various statistical properties of the time evolution and find a few general features that are characteristic for all MeSH hierarchies.

Although the empirical studies in this work are restricted to the networks between MeSH terms, it is quite plausible that a part of these features are more universal and occur in the time evolution of hierarchical networks in general.

1 Introduction

1.1 Problem statement

Hierarchical organization can be observed in large number of situations both in nature and society. Relevant research papers focusing on hierarchies mainly analyze real-world networks such as technological networks, ecological systems, social interactions and even neural networks.

Even though there are several different approaches for modelling the dynamics of real networks, there is no accepted method for describing the evolution of hierarchies specifically. The purpose of our work is to take a step towards a general model with a data-driven approach and analyze the hierarchical network of Medical Subject Headings (MeSH) provided by the NCBI (National Center for Biotechnology Information).

1.2 Contribution

By defining variables which describe the evolution process and creating several hierarchy attributes we reveal some significant evolution patterns that are characteristic for MeSH hierarchies, and our intuition is that a part of these recognized features might be universal and might occur in the time evolution of hierarchies in general.

We also partly contribute to solving the difficultness of maintaining the MeSH database. The maintenance of the thesaurus is quite challenging due to the large number of publications and the fast pace of the development of the biomedical field. The revealed patterns can help better understand the changes in MeSH.

Even though there are some other publications which aim to predict the expansion of MeSH terms, we reveal that the changes in the hierarchy not only consist of growth patterns but restructuring is also a significant feature. Our work gives a method for describing the whole phenomena of evolution thus contributing to the MeSH research with novel insights.

1.3 Structure of the report

Chapter 2 provides a summary of relevant papers mainly covering the addressed issues related to MeSH and shows the theoretical foundations of our analysis by introducing the concept of hierarchies, showing the relevance of network theory and introducing the results of the research works on the evolution and dynamics of networks. Chapter 3 presents our results and analysis of MeSH hierarchy starting from the descriptive analysis, identifying evolution patterns then the modelling process. Chapter 4 is the subjective part of our report where we interpret the results of Chapter 3, underscore the limitations of our work and expound some possible approaches for further improvement. Chapter 5 summarizes the results of our paper and briefly elaborates on further plans for analyzing the evolution of hierarchies.

2 Related work

2.1 MeSH

2.1.1 Data

MeSH is a hierarchy of terms called descriptors provided by the National Library of Medicine for indexing scientific publications and to facilitate detailed searching among them. These MeSH term databases are frequently updated and the yearly released files are available back to 1999 on PubMed. The sizes of these hierarchies are 1,000-10,000 nodes and organized into 16 different, yearly updated hierarchies such as Anatomy, Diseases, Chemicals and Drugs, etc. The natural representation of these hierarchies is given by directed acyclic graphs (DAGs), composed of links pointing from nodes higher in the hierarchy towards nodes in lower levels. Due to the appearance of new MeSH terms, the MeSH hierarchies are growing in time.

2.1.2 Addressed issues

There are numerous research works focusing on MeSH due to the large amount of publicly available data and the several problems caused by the difficultness of its maintenance. There are two main issues that require a comprehensive solution. The first one is the capability to annotate new scientific articles with MeSH terms effectively in a very accurate and rapid way. PubMed has already a quite fast methodology for it and there are already several proposed additional efficacious methods. For instance, [1] applies machine learning techniques for annotating new articles.

The other relevant issue is the maintenance of the thesaurus itself due to the large number of publications and the fast pace of the development of the biomedical field. There is no generally accepted efficient automated methodology for it and our work is addressing this problem and tries to reveal several patterns which can contribute to a better understanding of the behavior of MeSH evolution.

2.1.3 Related publications

There are three very related articles which address the evolution problem of biomedical vocabularies. The issue is generally analyzed under the field called ontology evolution [2].

The closest article to our works is [3], which applies machine learning methods to predict the expansion of MeSH terms based on temporal classifiers. They examined which MeSH terms the new descriptors will connect to and their analysis was based on several attributes: structural properties (depth, number of siblings, descendants of a term), citations (query results of the term and its descendants), annotations (number of assigned articles) and different combinations of them. They also analyzed temporal features and created classifiers from them (e.g. the acceleration of the increase in annotations). A list of MeSH terms of three hierarchies could be provided by their method where 80% of them expanded in the next period and they covered around half of the total expansion. Among the top 5 predictive attributes 3 of them were structural and temporal: the temporal sibling number, the temporal number of all descendants and the temporal number of direct descendants (children). The other 2 predictive variables were: the ratio of the annotation of all descendants and the number of descendants, and the number of query results of all descendants of a MeSH term.

Thus, [3] clearly indicates the importance of temporal classifiers and points out the difference compared to another similar research [4] where temporal features were not considered and the dataset was also different. It also emphasizes the complementary effect of the results with mentioning a sibling generation method proposed in [5] which is an approach how to suggest new siblings for a MeSH term based on text mining of structured HTML and published articles.

In our research we point out that the changes in the hierarchy are not only growth and expansion patterns which are described in [3] [4] [5] but also restructuring is a significant feature (when existing terms are continuously re-categorized). We aim to grasp the whole phenomena of the evolution thus contributing to the MeSH research with novel and unique insights.

There is a paper [6] which provide a general method for describing the evolution of a hierarchy by including not only addition but also deletion of terms, and creating a general similarity measure for quantifying the modifications. The research is constraint to one hierarchy (Psychology) and gives us some interesting examples and introduces a general insight about the magnitude of the changes in the MeSH dataset.

Our goal is to give not only a broad overview but also and in-depth analysis of the evolution of MeSH hierarchies. Furthermore, as [3] mentions MeSH is more like a hierarchy than an ontology (in contrast with the Gene Ontology analysis in [4]) thus our approach will focus on the hierarchy aspect.

2.2 Hierarchy

2.2.1 Definition

Hierarchical organization is a prevalent feature of many complex networks appearing in nature and society. Recent research works on hierarchy include transcriptional regulatory network of Escherichia coli [7], dominant–subordinate hierarchy among crayfish [8], leader–follower network of pigeon flocks [9], the rhesus macaque kingdoms [10], neural network [11], technological networks [12], social interactions [13] [14] [15], urban planning [16] [17], scientific journals [18], ecological systems [19] and evolution [20] [21].

Hierarchies can be divided into three main categories [18]:

- *Order hierarchy* which is based on a defined ranking, a partial ordering of the set of elements [22]
- *Nested hierarchy* (inclusion hierarchy or containment hierarchy) where items are aggregated into increasingly larger groups and the higher-level groups consist of smaller components which are more specific [23]

- *Flow hierarchy* which can be described as a directed graph, where the nodes are assigned to levels and lower-level nodes are influenced by the higher ones

2.2.2 Network Theory

Network theory can be an often-useful approach when we analyze hierarchies [24] [12] given the fact that most hierarchies can be modelled as a graph where the nodes represent the different elements while the directed edges indicate the hierarchical relationships between them.

Consequently, in case of MeSH hierarchies, which can be viewed as nested hierarchy due to the categorization of terms from general to more specific ones, network analysis is an advantageous approach too.

2.3 Evolution of Networks

The different approaches for modelling the evolution of real networks are summarized in detail in [25]. Most of the models are based on the scale-free property of real networks and introducing new concepts for preferential attachment and in some cases adding different attributes for nodes and edges such as fitness, aging, or inheritance behavior for better performance.

Most these models are either based on the degree of the nodes or on other newly added arbitrary parameters (e.g. age, fitness value). In our research, we are focusing on identifying the effect of different attributes in the evolution process.

3 Data Analysis and Results

3.1 Data Source

MeSH is a hierarchy of terms called descriptors provided by the National Library of Medicine for indexing scientific publications and to facilitate detailed searching among them. These MeSH term databases are frequently updated and the yearly released files are available back to 1999 on PubMed. The sizes of these hierarchies are 1,000-10,000 nodes and organized into 16 different, yearly updated hierarchies. In *Table 3.1* we can see all the hierarchies and their names. The datasets are available on different sites of PubMed [26] [27] with detailed explanation and several documentations [28] [29] [30].

A	Anatomy
B	Organisms
C	Diseases
D	Chemical and Drugs
E	Analytical, Diagnostic and Therapeutic Techniques and Equipment
F	Psychiatry and Psychology
G	Phenomena and Processes
H	Disciplines and Occupations
I	Anthropology, Education, Sociology and Social Phenomena
J	Technology, Industry, Agriculture
K	Humanities
L	Information Science
M	Named Groups
N	Health Care
V	Publication Characteristics
Z	Geographicals

Table 3.1 *Names of different hierarchies*

It is important to mention even though most of the MeSH terms appear in only one dataset (92%), there are some descriptors which belong to multiple hierarchies. The sizes of the hierarchies vary significantly and the largest ones are Organisms (B), Diseases (C) and Chemical and Drugs (D) as *Table 3.2* indicates. Our analysis is based on these three hierarchies to detect better the relatively small changes in the hierarchies throughout the years.

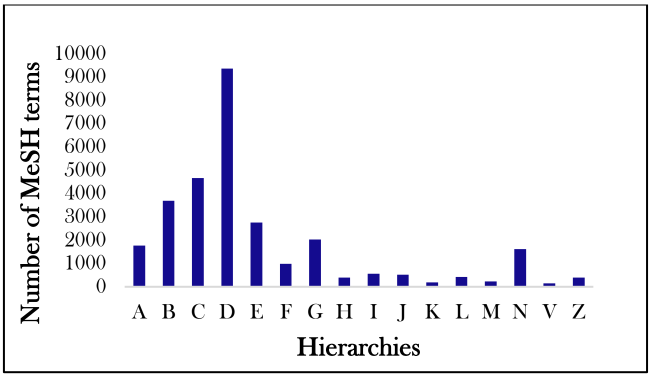

Table 3.2 *Number of MeSH terms in different hierarchies*

Most of the preprocessing and data cleaning has been already done as a part of my Project Laboratory work [31]. It mainly covered the data extraction from the XML format, the process of building up the network of the data and finally cleaning and selection of relevant date attributes. In *Table 3.3* the extracted attributes are listed, two of them are identifying the MeSH term while the Established Year attribute refers to the age of the descriptor which plays an important role in the evolution analysis.

Descriptor UI	Descriptor Unique Identifier. Seven-character alpha-numeric string uniquely identifying the record.
Descriptor Name	Name of the MeSH term
Date Established	First day of the first full month when the record first becomes available for searching in NLM's online databases, such as PubMed

Table 3.3 *Number of MeSH terms in different hierarchies*

3.1.1 Directed Acyclic Graph (DAG)

The hierarchies can be represented as directed acyclic graphs (DAGs) where the edges are directed to the lower levels and there is no cycle in the graph and there is no path to the higher levels from the lower ones. It is important to underscore that one MeSH can have not only multiple outgoing edges (children) but also multiple incoming edges (parents).

Figure 3.1 depicts a small part of a hierarchy and we can immediately infer the logical arrangement of the MeSH terms. For example, Ear is a child of Sense Organs and Head while the parents of Eyebrows are Hair and Eye as well. The sizes are proportional to the Total Degree of each node. For example, Musculoskeletal System and Body Regions are relatively large while Hip, Eyebrows and Eyelashes have few incoming and outgoing edges.

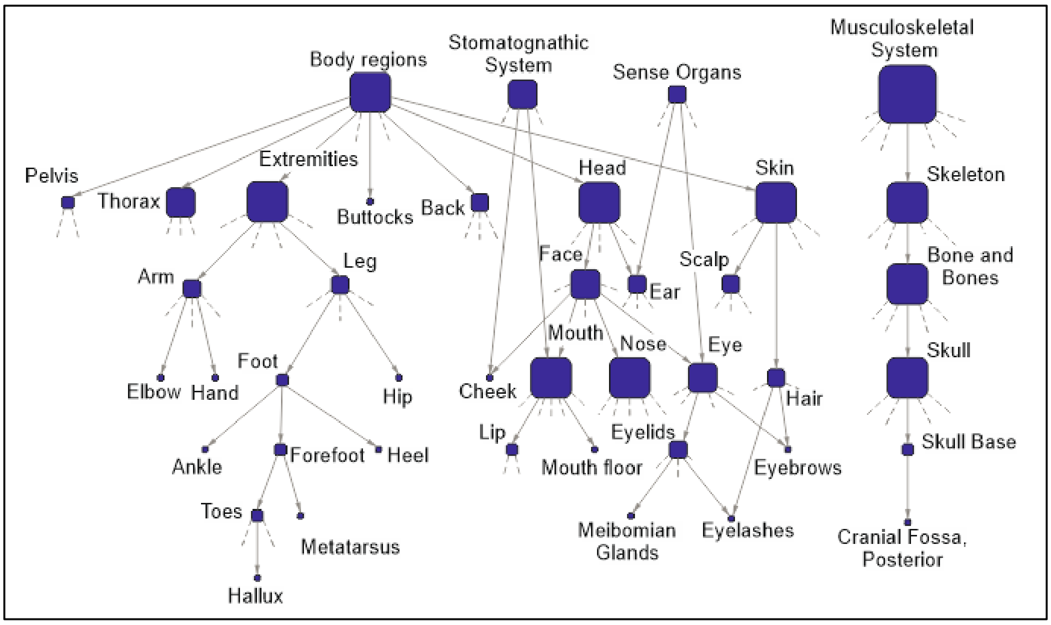

Figure 3.1 *Sample from the Anatomy hierarchy*

As [29] clearly describes "these trees should not be regarded as representing an authoritative subject classification system but rather as arrangements of descriptors for the guidance and convenience of persons who are assigning subject headings to documents or are searching for literature. Their structure frequently represents a compromise among the views and needs of particular disciplines and users, in the absence of any single universally accepted arrangement". So, there are several different approaches to arranging these MeSH terms in a logical manner and as we can see later the applied logic obviously changes over time.

3.1.2 Defining attributes

Based on the availability and the features of the datasets we created different types of attributes shown in *Table 3.4.* We could calculate all this attributes for all the MeSH terms in all hierarchies in 14 different years (from 2002 to 2015). The Annotation features [27] were only available from 2013.

Category	Name	Description
Time feature	Age	Subtracting the Established Year from the year we examined the MeSH term
Hierarchy features	Level (Maximum Depth)	Root nodes are on the first level, the longest path from them ranges from 1 to 12
	Minimum Depth	Minimum distance from the root, ranges from 1 to 12
	All Children Number	The number of nodes which are in the branch of the given MeSH term
	Sibling Number	The number of nodes which share the same parent
Network features	In Degree (Parent Number)	Number of incoming edges which is equal to the number of parents
	Out Degree (Children Number)	Number of outgoing edges which is equal to the number of children
	Total Degree	Number of total edges which is equal to the sum of incoming edges and outgoing edges.
Annotation features	Annotation	Number of articles assigned to as a Mesh term (with or without subheading)
	Major Annotation	Number of articles assigned to where marked as major MeSH term
	Annotation without subheading	Number of articles assigned to as a Mesh term where there was no subheading assigned
	Annotation with subheading	Number of articles assigned to as a Mesh term where there was subheading assigned

Table 3.4 Examined attributes of a MeSH term

3.1.3 Distributions of attributes

We created the distributions of the 12 defined attributes to understand our data better. The following figures are based on the 2016 version of Disease (C) hierarchy but most of the other hierarchies have similar features.

From *Figure 3.2* we can infer the shape of the hierarchy, most nodes are in the middle levels and their minimum distance from the roots are around 3 or 4 while there are relatively few top and bottom MeSH terms. The number of Levels and consequently the Minimum Depth is limited to 12 by NLM [29]. In *Figure 3.3* the distribution of the age of different nodes shows us that every year the number of new MeSH terms were similar except for the recent 10 years when the number of incoming nodes doubled.

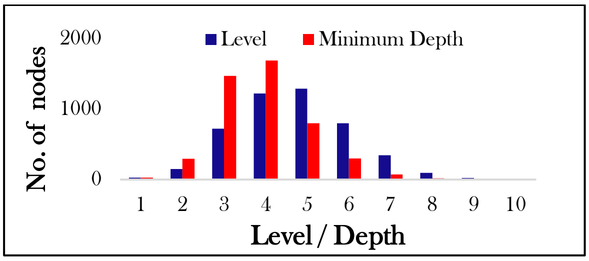

Figure 3.2 *Distribution of depths in the hierarchy of Diseases (C) in 2016*

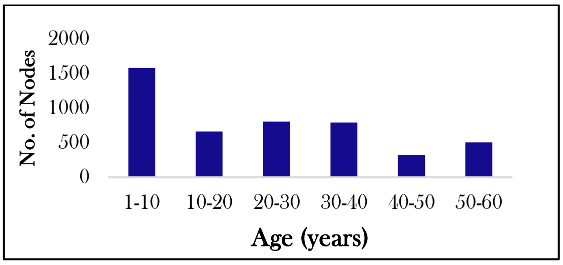

Figure 3.3 *Distribution of age in the hierarchy of Diseases (C) in 2016*

It is interesting that the degree distributions of the hierarchy follow a power law for in-degree and out-degree, thus the total degree, the sum of them shows a similar pattern (It is important to mention that around 3000 of the nodes are “leaves” and their out degree is 0). The power-law distributions can be observed in scale-free networks [25]. However, the absolute value of the slope of the fitted line is not between 2 and 3 as the scale-free model suggests, it is more than 10 in all cases.

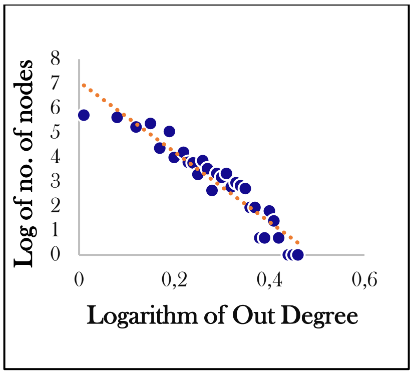

Figure 3.4 *Power law distribution of Out Degree for Disease hierarchy*

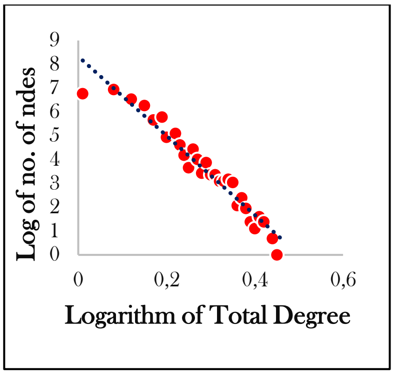

Figure 3.5 *Power law distribution of Total Degree for Disease hierarchy*

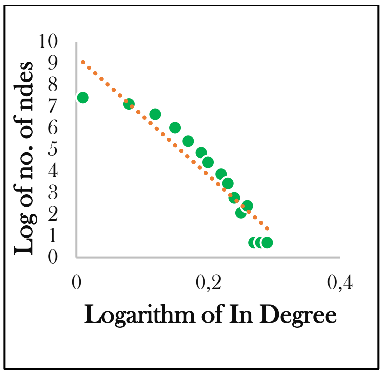

Figure 3.6 *Power law distribution of In Degree for Disease hierarchy*

In *Figure 3.5* we chose a logarithmic scale (natural) to show the distribution of the number of siblings and all children due to the large number of nodes which had very few siblings and terms in their branch as well.

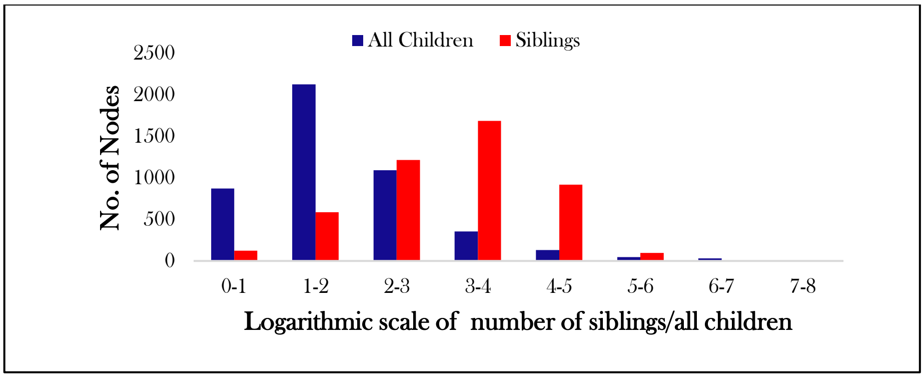

Figure 3.7 *Distribution of siblings and all children in the hierarchy of Diseases in 2016*

The distributions of the annotation numbers, which basically imply the popularity of each MeSH terms, can be seen in *Figure 3.6* and the logarithmic scale appeared to be advantageous in these cases too.

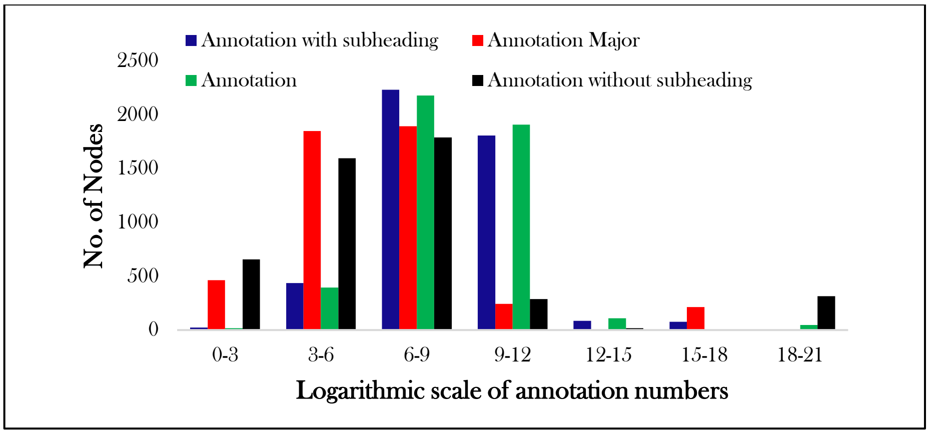

Figure 3.8 *Distribution of annotation numbers in the hierarchy of Diseases in 2016*

3.1.4 Correlation between attributes

By examining the scatter plots of each variables, then calculating the Pearson-correlation matrix (*Table 3.4*), we identified some relevant relationship properties between our attributes. We took the logarithm of most of the attributes to normalize their distributions, thus getting interpretable correlation values.

	Age	Log of All Children	Log of Total Degree	Log of In Degree	Log of Out Degree	Log of Sibling	Min Depth	Level	Log of Annot.	Log of Major Annot.	Log of Annot. (no s.)
Log of All Children	-0.004										
Log of Total Degree	-0.007	0.914									
Log of In Degree	-0.066	0.541	0.541								
Log of Out Degree	-0.036	0.973	0.973	0.342							
Log of SiblingNumber	-0.042	0.362	0.362	0.442	0.237						
Min Depth	-0.069	-0.226	-0.226	0.002	-0.128	-0.321					
Level	-0.079	0.031	0.031	0.439	0.009	-0.089	0.742				
Log of Annotation	0.121	0.261	0.261	-0.012	0.193	0.072	-0.202	-0.192			
Log of Major Annotation	-0.051	0.266	0.266	-0.097	0.204	0.043	-0.274	-0.295	0.717		
Log of Annot. (no subh.)	-0.166	0.278	0.278	0.066	0.207	0.126	-0.155	-0.125	0.738	0.669	
Log of Annot. (with subh.)	0.191	0.235	0.235	-0.044	0.185	0.048	-0.208	-0.201	0.959	0.691	0.691

Table 3.5 Pearson-correlation matrix of defined attributes

There is a significant correlation between All Children, Total Degree and Out Degree. The correlation of the last two is quite intuitive due to the much lower number of indegree compared to the outdegree of an average node.

Between Level (Maximum Depth) and Minimum Depth we can also realize a strong correlation effect. This can be explained with the fact that most of the nodes might have their parents from similar levels and not from completely different parts of the hierarchy.

High correlation values are prevalent between Annotation features too, especially between Annotation and Annotation with subheadings. This means that a MeSH Term with a high annotation number has high number for all attributes which belong to Annotation features.

3.2 Evolution patterns

3.2.1 Example

Figure 3.9 and *3.10* serve as an example for the evolution patterns in the MeSH hierarchies. The red edges and nodes are those which went under some change within this 1 year (from 2002 to 2003).

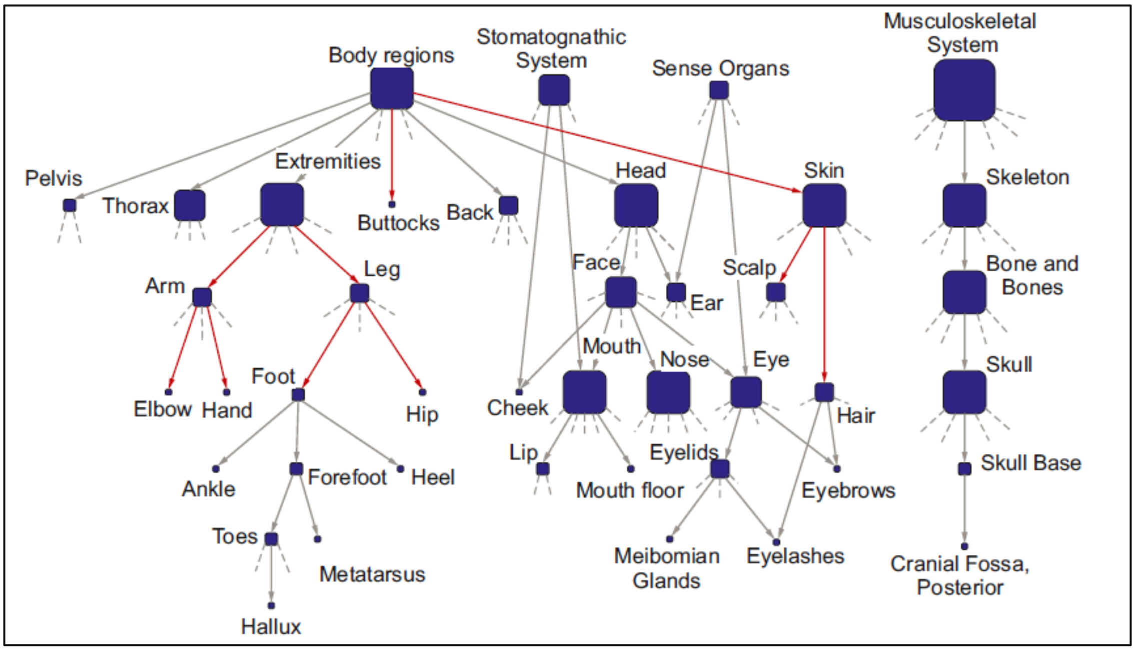

Figure 3.9 Sample from the Anatomy Hierarchy in 2002

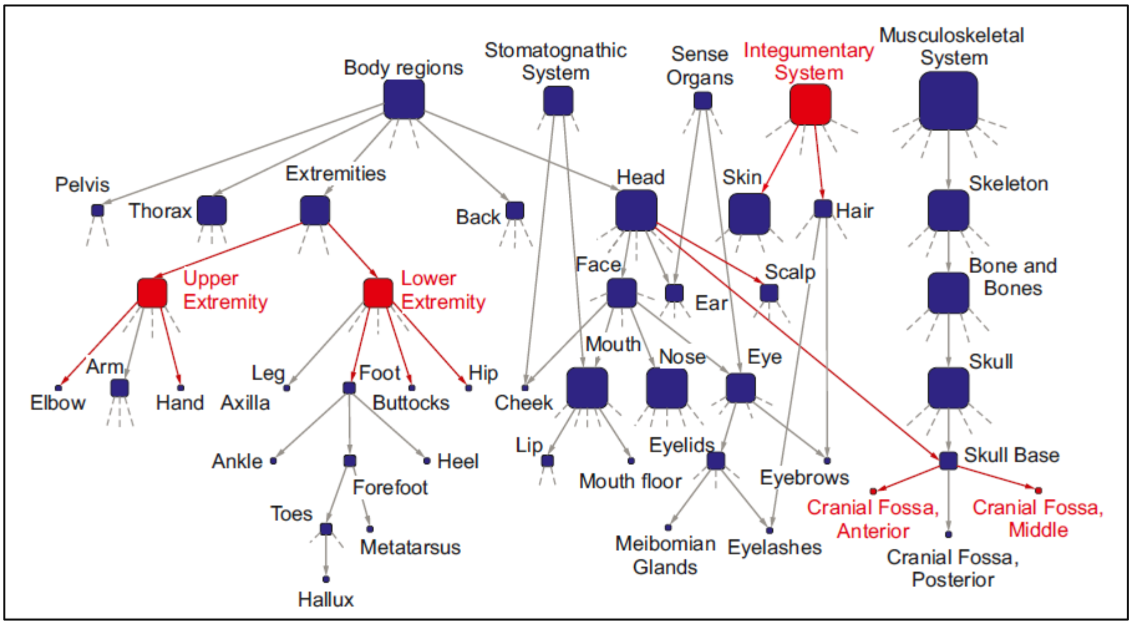

Figure 3.10 Sample from the Anatomy Hierarchy in 2003

As we can see the term Skin in 2002 had Body Region as its parent but in 2003 it was reorganized under Integumentary System which just joined the hierarchy. It also lost both Scalp and Hair as a child, because Hair went under the Integumentary System directly as well, becoming the sibling of its previous parent, and Scalp was connected to Head. Head got a new child, Skull Base from a completely different branch, and this Skull Base was expanded by two new children. On the other side of the hierarchy there was a significant rearrangement around the Extremities resulting a division to Upper Extremity and Lower Extremity branches. Foot and Hip was realigned under the newcomer Lower Extremity from Leg and Buttocks joined as a new sibling to them and left his previous position from being under Body Regions directly. Elbow and Hand went under similar changes and joined to Upper Extremity from their previous parent Arm.

In conclusion, the changes in the hierarchy are far from simple and obviously not only consists of expanding branch features, several rearrangements can also be detected.

3.2.2 Defining growing and restructuring

The purpose of our work is to describe the whole evolution process and to not only focus on the expansion. Our current approach (after several previous ones) is based on the edges and specifically on the direction of them as *Table 3.5* indicates.

Growing event	The number of new outgoing edges of a MeSH Term *The number of Growing events is x for a MeSH term being part of a hierarchy in year n, if in year n+1 MeSH is still part of the hierarchy and got x new outgoing edges (we do not count with the deletions, so it is not the absolute increase of outdegree)*
Restructuring event	The number of new incoming edges of a MeSH Term *The number of Restructuring events is x for a MeSH term being part of a hierarchy in year n, if in year n+1 MeSH is still part of the hierarchy and got x new incoming edges (we do not count with the deletions, so it is not the absolute increase of indegree)*

Table 3.6 Definition of evolution patterns

This way we can easily separate expansion and rearrangement modifications in the hierarchy and we can also make it quantifiable for each MeSH term. It is important to mention that we do not take the degree changes of completely new nodes into account during our calculations and we also neglect the deletions. We also do not differentiate between the cases when there are new nodes on the other end of the new edges or old ones.

We can measure these changes yearly given the rate of the updates of the hierarchies by PubMed. *Figure 3.10* shows the yearly proportion of Growing and Restructuring events in case of Disease hierarchy and we can see that the ratio of Restructuring events is not negligible varying from around 10% to 50% of the total changes.

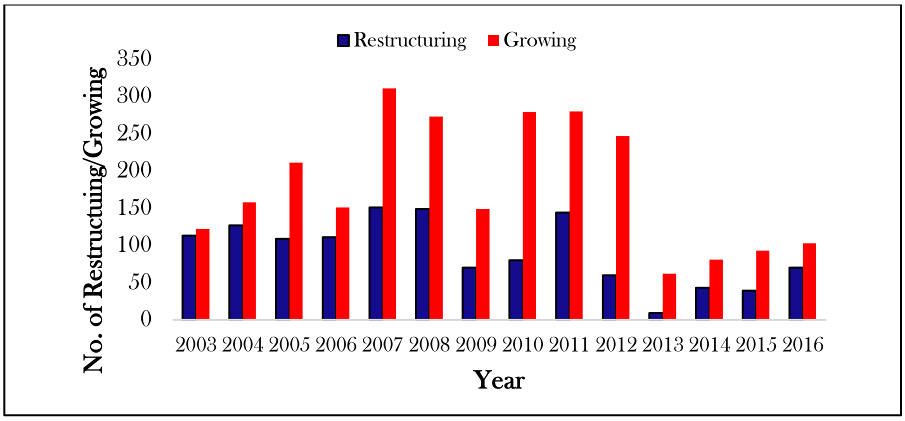

Figure 3.11 Growing and Restructuring for the Disease (C) hierarchy from 2003 to 2016

In *Figure 3.12* and *3.13* the distribution of Growing events and Restructuring can be seen, and both follow a power law, similarly to the previously described degree distributions (*Figure 3.4, 3.5* and *3.6*). This means for instance that the number of nodes with very few Growing events is significantly large, while nodes with relatively high number of new outgoing edges are quite scarce.

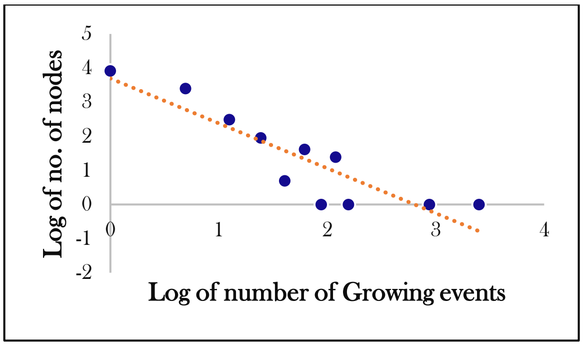

Figure 3.12 Power law distribution of growing events for the Disease (C) hierarchy from in 2016

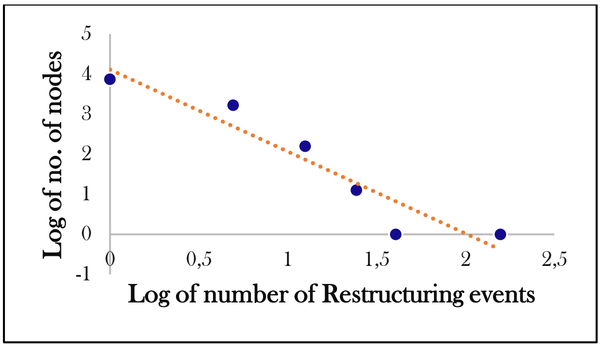

Figure 3.13 Power law distribution of restructuring events for the Disease (C) hierarchy from in 2016

The charts above do not contain the 0 values; thus, we should highlight the relatively small number of occurrences of Growing and Restructuring events compared to the size of the hierarchies. The proportion of the nodes which get new incoming or outgoing edges is usually less than 4 percent in each year for most datasets.

	A	B	C	D	E	F	G
Average number of nodes with Growing events	1,9%	1,4%	1,9%	1,7%	2,5%	2,0%	3,0%
Average number of nodes with Restructuring events	1,5%	1,0%	1,2%	1,5%	1,7%	1,1%	3,9%

Table 3.7 Average ratio of effected nodes by Growing and Restructuring (yearly) for the largest hierarchies

This low ratio can cause several issues during modelling, mainly leading to predictive models with low recall rate and high precision due to the imbalanced dataset feature.

3.2.3 Parent-Child relationships

Age

As we could see earlier the age distribution of MeSH terms is quite flat except for the last 10 years (*Figure 3.3*), but it shows some correlation with level and minimum depth. In *Table 3.7* the proportions show the ratios when the target node or the source node is younger, aggregated for all edges in each hierarchy.

Younger source	Younger target	Hierarchy
8%	92%	K. Humanities
16%	84%	L. Information Science
18%	82%	N. Health Care
19%	81%	D. Chemical and Drugs
20%	80%	B. Organisms
22%	78%	F. Psychiatry and Psychology
24%	76%	I. Anthropology, Education, Sociology and Social Phenomena
25%	75%	H. Disciplines and Occupations
27%	73%	J. Technology, Industry, Agriculture
28%	72%	E. Analytical, Diagnostic and Therapeutic Techniques and Equipment
30%	70%	A. Anatomy
31%	69%	Z. Geographicals
32%	68%	M. Named Groups
32%	68%	C. Diseases
37%	63%	G. Phenomena and Processes
65%	35%	V. Publication Characteristics

Table 3.7 Comparing the proportion of younger target and older source nodes for all edges in each hierarchy

In almost all hierarchies (except for Publication Characteristics) we can see that usually the source node is the older one. It means that generally the lower-level nodes are younger, but there are some younger nodes which attach to higher levels. Or it can be also interpreted as an upward wandering of nodes in the hierarchy and older nodes soon become source nodes rather than target ones.

Annotations

In case of annotations a very similar feature can be observed, in most cases, the higher annotation numbers belong to the source node and the lower ones to the target nodes. It is true for all annotation features (Annotation, Major Annotation, Annotation without subheading, Annotation with subheading) and the ratio varies between 60% to 80% in favor of the source nodes.

3.2.4 Evolution effects of different attributes

The following figures (*Figure 3.14 – 3.23*) were made by an aggregating and binning method to identify the major influencer attributes in the evolution process. Due to the low number of incoming and outgoing edges, the general features of the hierarchies do change significantly throughout the 13 years that we analyzed. Consequently, the distributions (*Figure 3.2-3.8*) can be considered static and we can analyze each attribute effect by taking the average of Growing and Restructuring events aggregated on the distributions (most of these figures have log-log axis) neglecting the dimension of time. It is important to underline that the if we take the attributes of a MeSH term from year *n* then we calculate the number of Growing or Restructuring events that effected it by comparing year *n* and year *n+1.*

For example, in Figure 3.14 the horizontal axis, the logarithm of In Degree, is divided into 20 equal range bins and the first data point shows that the average number of Growing events for nodes with relatively low In Degree in the previous year (for those who are inside the bin) is quite low (this aggregation covers every year).

The same methodology applies to the other graphs too, all of them have a horizontal axis with 20 bins which serve as a basis for calculating the averages.

Preferential attachment, network features

Scale-free networks with power law distributions can be observed in real systems, and preferential attachment model is an appropriate way of describing of their patterns. In [25] preferential attachment is proved for some real networks.

In our case, we are not exactly examining the same situation due to the definition of Growing and Restructuring. We mainly observe the attachment of edges, not nodes, we have a directed network and indeed we separate the attachment of outgoing and incoming edges.

Furthermore, our network is a special one, a directed acyclic graph, and we could also see from the degree distribution that despite the power law it is significantly different from scale-free networks which was shown by the steep slope of the fitted line (in *Figure 3.4-3.6*).

In *Figure 3.14* and *3.15* a special kind of preferential attachment can be seen, the higher the In Degree the more Restructuring events happens on average. This feature can be visible for most of the hierarchies, two examples illustrate the similarity between them.

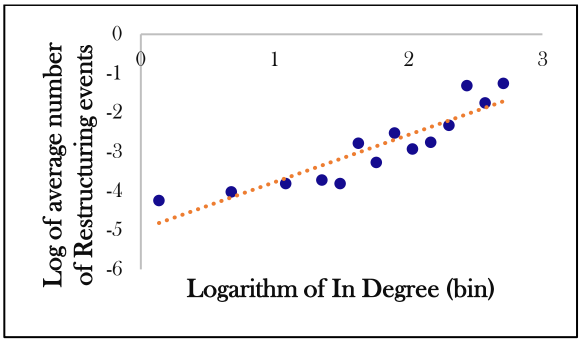

Figure 3.14 Average number of Restructuring events for binned In Degree for the Disease (C) hierarchy

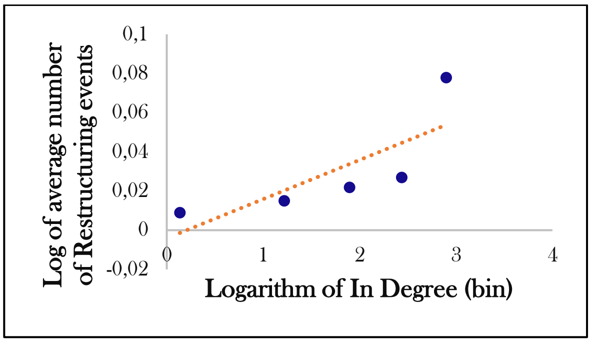

Figure 3.15 Average number of Restructuring events for binned In Degree for Organisms (B) hierarchy

In *Figure 3.16* and *3.17* another special kind of preferential attachment can be observed, the higher the Out Degree the more Growing events happens on average. This feature can be seen in most cases, here the Disease and the Chemical and Drugs hierarchies serve as illustrations. (The effect of Total Degree is the same as the impact of Out Degree due to the high correlation)

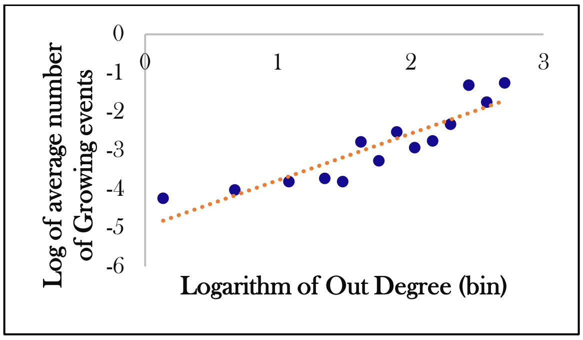

Figure 3.16 Average number of Growing events for binned Out Degree for the Disease (C) hierarchy

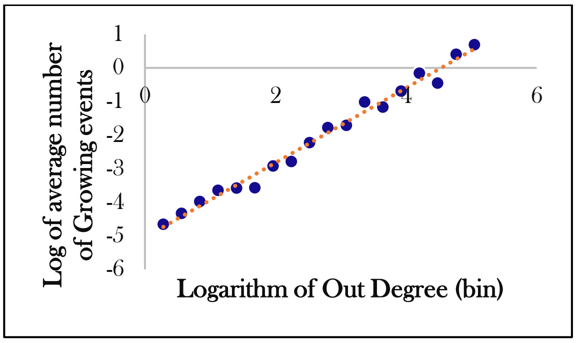

Figure 3.17 Average number of Growing events for binned Out Degree for Chemical and Drugs (D) hierarchy

These two features are very similar to preferential attachment. The probability of getting new outgoing edges is bigger when the node already has higher number of existing ones and the probability of getting new incoming edges is also bigger when the node already has higher number of parents.

Hierarchical features

Figure 3.18 and *3.19* show that higher the number of siblings the more Restructuring events happens on average. This feature can be visible for most of the hierarchies, two examples illustrate the similarity between them. Here we should mention that this pattern is not always the same for all hierarchies (e.g. for the Disease (C) hierarchy this trend does not exist).

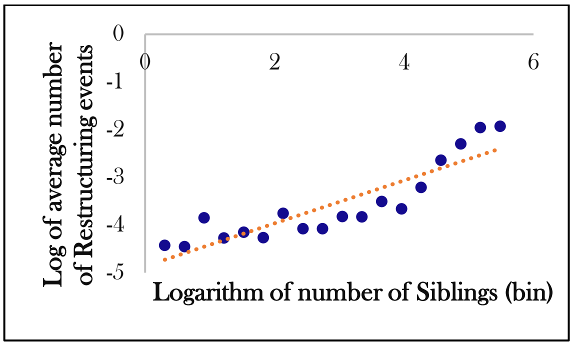

Figure 3.18 Average number of Restructuring events for binned Singling number for the Disease (B) hierarchy

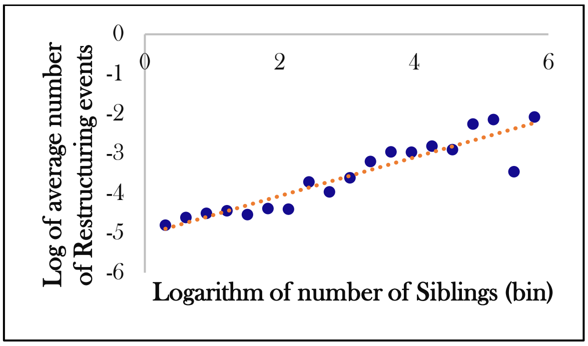

Figure 3.19 Average number of Restructuring events for binned Singling number for Chemical and Drugs (D) hierarchy

The previously described pattern might seem logical if we think about the probable effect when a node has more siblings. One would think that the probability of their arrangement of these terms can certainly depend on their complexity which can also be measured with the number of siblings.

Figure 3.20 and *3.21* refer to a pattern that higher the number of nodes in the sub branch of a node the more Growing events effects it on average. This feature is true for most Mesh datasets; two of them can be observed here.

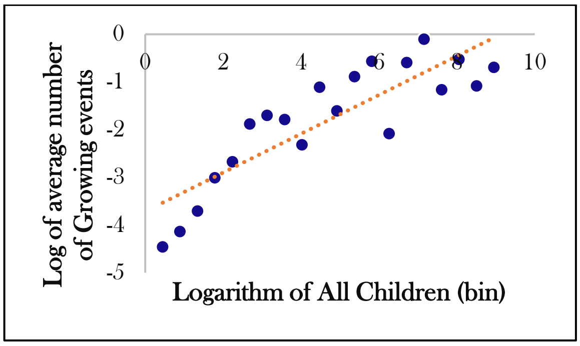

Figure 3.20 Average number of Growing events for binned In Degree for the Organisms (B) hierarchy

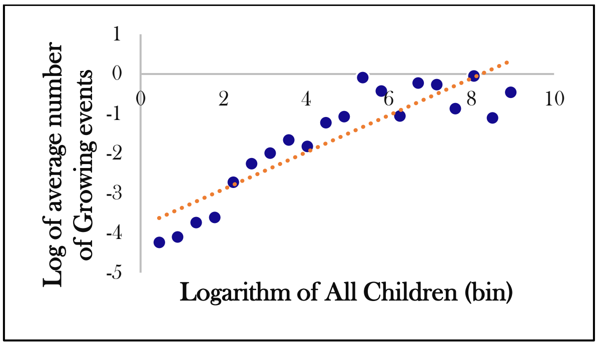

Figure 3.21 Average number of Growing events for binned All Children for Chemical and Drugs (D) hierarchy

This feature might not be that intuitive as the previous ones. But due to the high correlation between All Children and Out Degree this result is also understandable.

In *Figure 3.22* and *3.23* we can see two hierarchies when top level nodes usually get more outgoing edges on average than the lower ones (in log-log scale the scatter plot could be also fitted to a straight line but it is not that meaningful to take the logarithm of the discrete values of Levels, and here we did not have to bin twice, the separate levels served as bins).

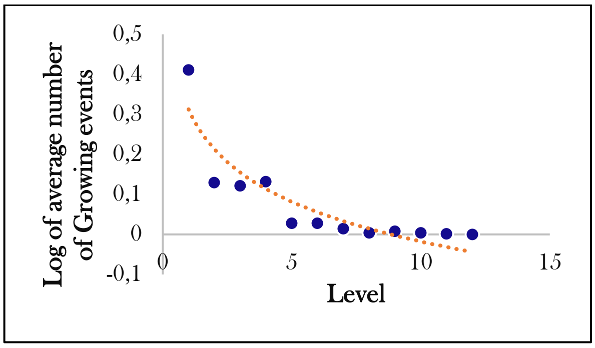

Figure 3.22 Average number of Growing events for Level number for the Organisms (B) hierarchy

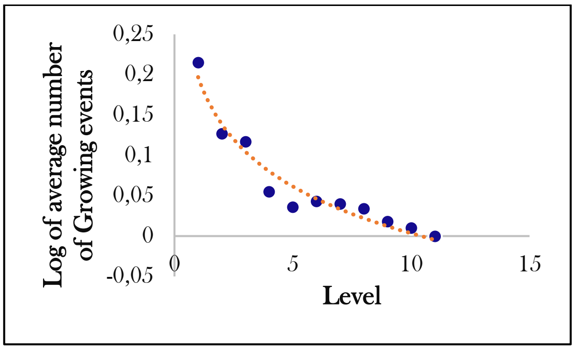

Figure 3.23 Average number of Growing events for Level number for Chemical and Drugs (D) hierarchy

In case of Minimum Depth, the effect is similar (owing to the strong correlation between them). This result may be a little against our intuition that a hierarchy evolve mainly by extending its lower levels. But we have to pay attention that our model does not only consider the expansions by new MeSH terms but also reconnections to existing ones when we talk about new outgoing edges.

Annotation features

Figure 3.24 and *3.25* show that higher the number of Major Annotations the more Growing events happens on average for most of the hierarchies, the Organisms (B) and Chemical and Drugs (D) hierarchies are our illustrations this time.

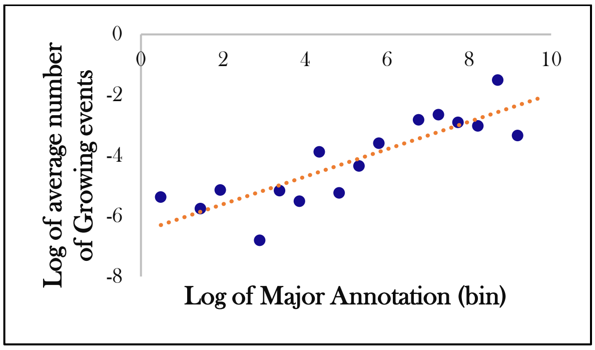

Figure 3.24 Average number of Growing events for binned Major Annotation number for the Organisms (B) hierarchy

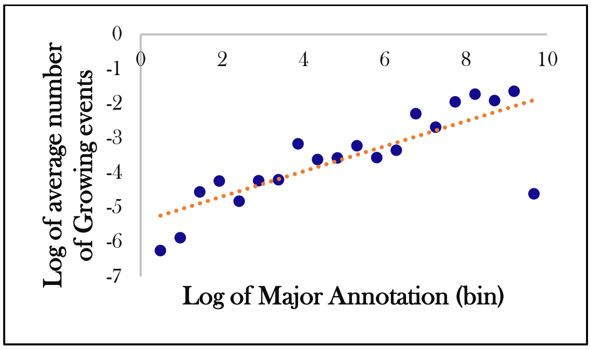

Figure 3.25 Average number of Growing events for binned Major Annotation number for Chemical and Drugs (D) hierarchy

The Annotation feature can be construed as a popularity measure for each MeSH term, the more popular a descriptor is the more articles it gets assigned to. This interpretation makes the results above more comprehensible and straightforward, a popular node has a higher chance to expand.

Other Annotation features, such as Annotations without subheading, due to the correlation between each other, also indicate analogous formula.

Summary of the effects

In conclusion, these observations revealed strong relationships between the defined attributes and Growing and Restructuring event numbers. For Restructuring the identified attributes were In

Degree (preferential attachment) and Sibling Number while for Growing we found Level, Total Degree, Out Degree (preferential attachment), All Children and most Annotation features relevant.

3.3 Modelling

3.3.1 Question

Our purpose is to predict which MeSH term is effected by Growing or Restructuring event in the next year. Thus, the target variables (*Table 3.8*) are not the number of Growing events or Restructuring events for each MeSH term, it is a binary value for both cases. The value is 1 if the number of events exceeds 0, and it 0 in the rest of the cases.

Target variable	Possible values
Growing Binary	0/1
Restructuring Binary	0/1

Table 3.8 Target variables

3.3.2 Input variables

We defined the input variables based on the previously described attributes. It is important to underline that the we take the attributes of a MeSH term from year *n* and we aim to predict the value of Growing Binary and Restructuring Binary in year *n+1.*

Category	Name
Time feature	Age
Hierarchy features	Level (Maximum Depth)
	Minimum Depth
	All Children Number
	Sibling Number
Network features	In Degree (Parent Number)
	Out Degree (Children Number)
	Total Degree
Annotation features	Annotation
	Major Annotation
	Annotation without subheading
	Annotation with subheading
Logarithm of all attributes	**Log*attribute name***
Temporal attributes	**Temp*attribute name* =** $\frac{\boldsymbol{attribute(in\ year\ n) - attribute(in\ year\ n-1)}}{\boldsymbol{attribute(in\ year\ n)}}$

Table 3.9 Input variables

3.3.3 Method and model

Training and testing dataset

In our predictive model training dataset was all the data before 2016, while the testing data was the records that belonged to 2016.

Imbalanced data

Based on our previous descriptive analysis we face an imbalanced data issue due to the low ratios of Growing and Restructuring events (*Table 3.7*). There are several research works which aim to find a general solution for it, and mostly they recommend random undersampling or random oversampling [32] [33]. Based on the results and recommendations we chose the random undersampling due to the over-fitting problem of random over sampling [32].

During our modeling, we balanced the data to different ratios to find the best proportion: 5%-95%, 10%-90%, 30%-50% and 50%-50%. We got the best result for 10%-90% balance (*Figure 3.26*). (The explanation of F-measure and the model can be read in the next section)

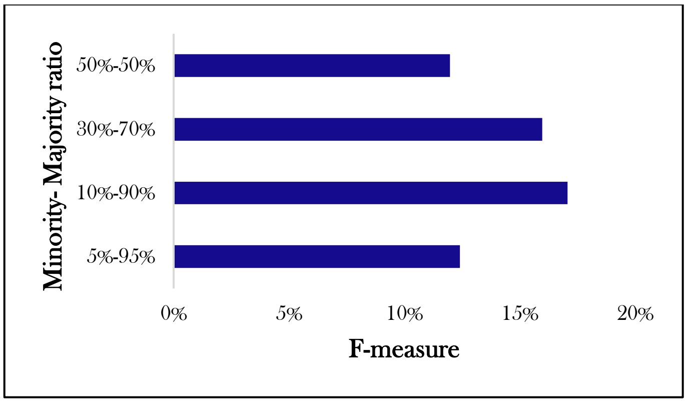

Figure 3.26 Random under sampling results for the predictive model of hierarchy E

Selecting model

We did not only use the built-in model recommendation by SPSS Modeler, but also manually compared the performance of them, and finally selected the CHAID (Chi-squared Automatic Interaction Detection) model.

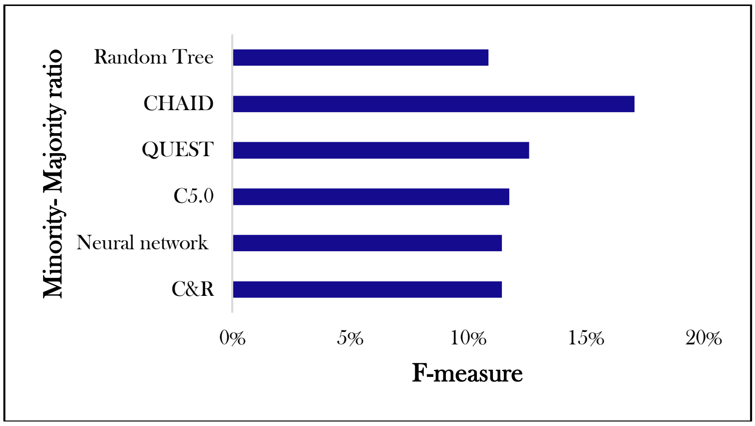

Figure 3.26 Results of different models for hierarchy E

3.3.4 Evaluation metrics

Due to the still quite imbalanced data despite the random undersampling we cannot use accuracy for measuring how good our model. Therefore, we select F-measure which weights recall and precision the same way. The definition of F-measure is based on the Recall and Precision which are based on the confusion matrix (*Figure 3.27*)

		Prediction outcome	
		1	0
Actual value	1	True Positive	False Negative
	0	False Positive	True Negative

Table 3.10 Confusion matrix

The definitions of Recall and Precision are the following:

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \qquad Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

Equation 3.1 Definition of Recall and Precision

The definition of F-measure is the harmonic mean of Recall and Precision:

$$F - measure = \frac{2 * Recall * Precision}{Recall + Precision}$$

Equation 3.2 Definition of F-measure

3.3.5 Results

We calculated the value of the confusion matrix, precision, recall and F-measure manually based on the result of the testing data. We show the main outputs for the hierarchy E (Analytical, Diagnostic and Therapeutic Techniques and Equipment).

Growing

The confusion matrix results:

		Prediction outcome	
		1	0
Actual value	1	9	53
	0	14	2684

Table 3.11 Confusion matrix for modelling Growing for hierarchy E

Precision, Recall and F-measure:

$$Recall = 14{,}52\% \qquad Precision = 39{,}13\% \qquad F-measure\ = 21{,}18\%$$

Equation 3.3 Results of precision, recall and F-measure for modelling Growing for hierarchy E

Predictor importance produced by the model:

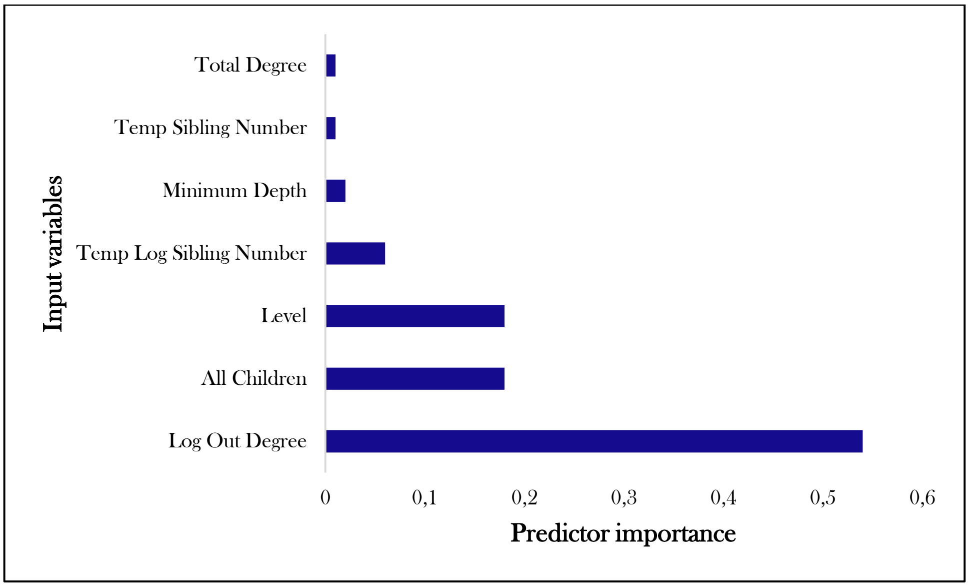

Figure 3.27 Predictor importance of input variables for modelling Growing for hierarchy E

Restructuring

The confusion matrix results:

		Prediction outcome	
		1	0
Actual value	1	19	15
	0	329	2397

Table 3.12 Confusion matrix for modelling Restructuring for hierarchy E

Precision, Recall and F-measure:

$$Recall = 55{,}88\% \qquad Precision = 5{,}46\% \qquad F - measure\ = 9{,}95\%$$

Equation 3.4 Results of precision, recall and F-measure for modelling Restructuring for hierarchy E

Predictor importance produced by the model:

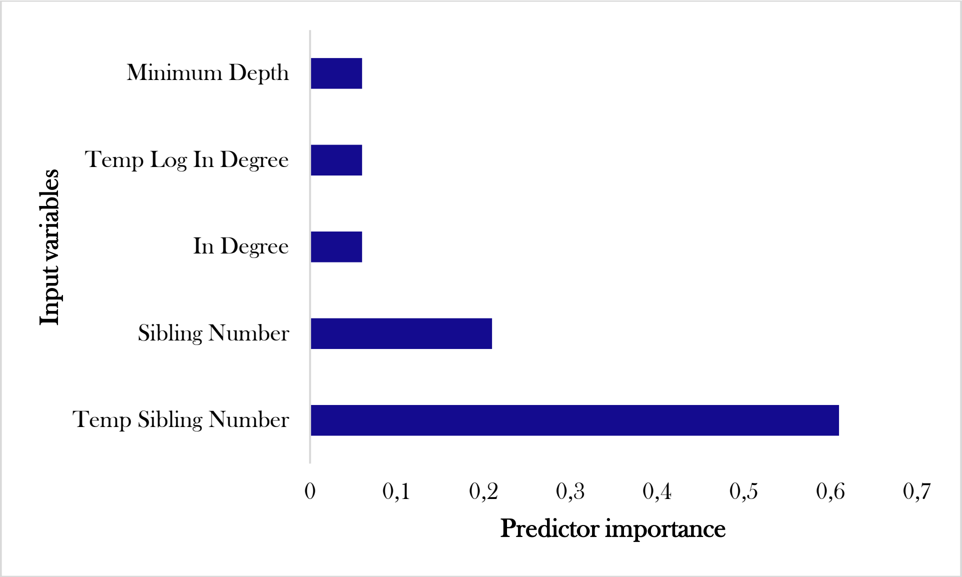

Figure 3.28 Predictor importance of input variables for modelling Restructuring for hierarchy E

Evaluation of the model results

Even though the F-measures are not very high in both Growing and Restructuring, we can see that it successfully validated our previously described evolution patterns.

For Restructuring it also identified Sibling Number and In Degree as an important attribute and for Growing it also found Level, Out Degree, All Children relevant. One more takeaway from the model results is that the temporal attributes also play an important role in the evolution of hierarchies.

4 Discussion

4.1 Interpreting the results, limitations

4.1.1 Definition, approach

Our results are strictly based on our current approach, the definition of Growing and Restructuring. Even though this method is simple and intuitive, we neglect the deletions and we also do not differentiate between the cases when there are new nodes on the other end of the new edges or old ones which might suggest different scenarios.

One logical approach would be a completely edge based one where we define the evolution patterns with *Table 4.1* considering the fact whether the ends of the new edge are connected to an old node or a new node. This way we can model rewiring and deletion as well, while group attachments remain intangible.

Target \ Source	Old	New
Old	Rewiring / Deletion	Growing
New	Restructuring	Group attachment

Table 4.1 A different, edge based approach

We can realize that the previously mentioned model conflicts with our current method because those edges which are rewired are included in the definition of Growing and Restructuring, basically we count them twice.

Target \ Source	Old	New
Old	Growing and Restructuring	Growing
New	Restructuring	Group attachment

Table 4.2 Our current method in the perspective of another model

This difference does not necessary mean that one of the approach is not correct, further research can tell which model works better.

4.1.2 Defining attributes

In our research, we defined several different attributes and we also divided them into different categories (Time, Hierarchy, Network and Annotation feature). We considered these ones as the most important ones but it might be probable that we did not include some other relevant ones. There are several network properties (clustering coefficient, centrality measures, distance, etc.) which are mostly irrelevant due to the hierarchical feature of our network. There are also global characteristics (density, average path length, etc.) which are also not important for us since we based our analysis on the stability of the whole hierarchy (low number of changes), so these global variables are constant too.

However, a more important fact is that we only considered one year data for describing the evolution patterns of the following year (e.g. the acceleration and history of different attributes might play some role too), and we only included them in our predictive models. In [3] the importance of temporal classifiers is emphasized and among their top 5 predictive attributes 3 of them were not only structural but also temporal: the temporal sibling number, the temporal number of all descendants and the temporal number of direct descendants (children). Furthermore, one of their other predictive variables were the number of query results of all descendants of a MeSH term which we did not consider at all.

Even though that our approach is completely different from [3], temporal features or other external attributes might be relevant in our case as well.

4.1.3 Dataset

Our results and analysis is limited to this specific dataset; we cannot derive a general conclusion about the evolution behavior of other hierarchies. MeSH has some features which might not be prevalent in other cases, one of it is the limited number of levels (12) which might influence significantly the changes in the dataset (even though most hierarchies has not reached this maximum number yet).

The examination of the evolution is also limited due to the frequency of updates (yearly) and the amount of publicly available data (form 2002 – 2016). On the other hand, each MeSH term has an Established Date assigned which not only consists of year but also, month and day. In our research, we did not analyze the month and day value, because the structural changes are only recorded yearly.

Another relevant feature of the data is that is artificially generated and maintained. MeSH Terms are added and updated manually and our observations might not be valid for those hierarchies where the construction of the data is less centralized.

4.2 Further improvements

Based on our analysis in the chapter of Limitations, our work might be further improved by creating broader definitions for describing the evolution process and finding some relevant additional attributes might be also advantageous.

5 Conclusion

5.1 Results

The main results of our work are the identified evolution patterns and the selection of significant attributes which effect the changes in the hierarchies over time. Furthermore, we also depicted how these attributes contribute to the evolution process.

To achieve this outcome, we defined two variables which describe the evolution process – Growing and Restructuring – and created several descriptive attributes for each MeSH term. While our methods and definitions for evolution could be further enhanced, the current results not only show that the evolution is not random but also indicate that there are numerous attributes which contribute significantly to the process.

For Restructuring the identified influential attributes were In Degree and Sibling Number while for Growing we found Level, Total Degree, Out Degree, All Children and most Annotation features relevant *(Figure 3.14-3.25)*. We could observe a preferential attachment behavior due to the power law correlations between Restructuring and In Degree, and between Growing and Out Degree.

We showed that creating a predictive model based on the defined attributes could be a promising prospect despite the imbalanced data issue. Our initial predictive models validated our recognized patterns and suggested that not only static but also temporal attributes play an important role in evolution for both Restructuring and Growing.

Our research work can help better understand the changes in MeSH, thus complementary to the research works focusing on solving the difficultness of maintaining the MeSH database. We also hope that it can also serve as a basis for further research on the evolution of hierarchies.

5.2 Future work

We aim to progress in the current direction since the results are promising, and try to reveal some other interesting patterns and finalize the predictive modelling part. Based on our result analysis in the discussion section we may extend our definitions and try to identify more attributes which are also relevant.

A fascinating question is whether our revealed patterns are ubiquitous and applies to the evolution of hierarchies in general, so a certain next step is to validate the concept on other datasets. For instance, one interesting dataset could be an evolving hierarchical brain network (e.g. in [34] the researchers did not only succeed to model the brain data with directed edges but also revealed an interesting evolution behavior).

Acknowledgements

I would like to express sincere gratitude and appreciation to *Dr. Palla Gergely* (MTA-ELTE Statistical and Biological Physics Research Group) and *Barta Gergő* (Department of Telecommunications and Informatics) for their advice and support.

I would also like to thank *Zagyva Tamás* for sharing his previous conference experiences, his constant support and countless beneficial tips.

List of Figures and Tables

References

[1] G. M. N. T. S. D. H. S. M. Tsatsaronis, "A maximum-entropy approach for accurate document annotation in the biomedical domain," *BMC Journal of Biomedical Semantics,* vol. 3, no. 1, 2012.

[2] P. M. T. Leenheer, "Ontology evolution," *Ontology Management,* p. 131–176, 2008.

[3] G. V. I. K. N. &. N. K. Tsatsaronis, "Temporal classifiers for predicting the expansion of medical subject headings," *International Conference on Intelligent Text Processing and Computational Linguistics,* pp. 98-113, 2013.

[4] C. C. F. Pesquita, "Predicting the extension of biomedical ontologies," *PLoS Computational Biology,* vol. 8, no. 9.

[5] G. W. T. S. M. Fabian, "Extending ontologies by finding siblings using set," *Bioinformatics,* vol. 28, no. 12, pp. 292-300, 2012.

[6] A. T. a. K. L. McCray, "Taxonomic change as a reflection of progress in a scientific discipline," *Evolution of Semantic Systems,* pp. 189-208, 2013.

[7] H. W. B. J. &. Z. A. P. Ma, "Hierarchical sructure and modules in the Escherichia coli transcriptional regulatory network revealed by a new top-down approach," *BMC Bioinformatics,* vol. 5, p. 199, 2004.

[8] H. C. a. H. R. Goessmann C, "The formation and maintanance of crayfish hierarchies: Behavioral and self-structuring properties.," *Behavioral Ecology and Sociobiology,* vol. 48, pp. 418-428, 2000.

[9] Á. Z. B. D. a. V. T. Nagy M, "Hierarchical group dynamics in pigeon flocks," *Nature,* vol. 464, p. 890–893, 2010.

[10] M. M. B. B. a. M. B. Fushing H, "Ranking network of captive rhesus macaque society: A sophisticated corporative kingdom.," *PLoS ONE,* vol. 6, p. e17817, 2011.

[11] M. H. C. C. &. K. R. Kaiser, "Hierarchy and dynamics of neural networks," *Front. Neuroinform,* vol. 4, p. 112, 2010.

[12] P. D, Hierarchy in Natural and Social Sciences, Volume 3 of Methodos Series., Dodrecht, the Netherlands: Springer, 2006.

[13] D. L. D.-G. A. G. F. a. A. A. Guimerà R, "Self-similar community structure in a network of human interactions," *Physical Review E,* vol. 68 , p. 065103, 2003.

[14] P. G. a. V. T. Pollner P, "Preferential attachment of communities: The same principle, but a higher level," *Europhysics Letters,* vol. 73, p. 478–484, 2006.

[15] V. S. a. S. RV, "Self-organization versus hierarchy in open-source social networks," *Physical Review E,* vol. 76, p. 046118, 2007.

[16] B. M. n. L. P, Fractal Cities: A Geometry of Form and Function, San Diego, CA, 1994.

[17] P. R. Krugman, "Confronting the mystery of urban hierarchy," *J. Jpn. Int. Econ.,* vol. 10, pp. 399-418, 1996.

[18] G. T. G. M. E. P. P. &. V. T. Palla, "Hierarchical networks of scientific journals," vol. 1, 2015.

[19] H. H. a. U. R, "Information theoretical analysis of the aggregation and hierarchical structure of ecological networks," *Journal of Theoretical Biology,* vol. 116, p. 321–341.

[20] E. N, Unfinished Synthesis: Biological Hierarchies and Modern Evolutionary Though, Oxford University Press: New York, 1985.

[21] M. DW, "The hierarchical structure of organisms," *Paleobiology,* p. 405–423, 2001.

[22] L. D, Hierarchy, Complexity, Society, Dordrecht, the Netherlands: Springer, 2006.

[23] W. E. T, Nested Ecology: The Place of Humans in the Ecological Hierarchy, Baltimore, MD.: John Hopkins University Press, 2009.

[24] S. A. M. D. O. Z. a. B. A.-L. Ravasz E, "Hierarchical organization of modularity in metabolic networks," *Science,* vol. 297, p. 1551–1555, 2002.

[25] R. a. A.-L. B. Albert, "Statistical mechanics of complex networks," *Reviews of modern physics,* vol. 74, no. 1, p. 47, 2002.

[26] "National Library of Medicine, MeSH database," 2016. [Online]. Available: https://www.nlm.nih.gov/mesh/filelist.html.

[27] MEDLINE/PubMed, "MEDLINE/PubMed Baseline Repository (MBR)," [Online]. Available: https://mbr.nlm.nih.gov/Downloads.shtml#Hist.

[28] "National Library of Medicine, Medical Subject Headings Fact Sheet," 2016. [Online]. Available: https://www.nlm.nih.gov/pubs/factsheets/mesh.html.

[29] "National Library of Medicine, MeSH Tree Structures," 2016. [Online]. Available: https://www.nlm.nih.gov/mesh/intro_trees.html.

[30] "National Library of Medicine, MeSH XML Data ELements," 2016. [Online]. Available: https://www.nlm.nih.gov/mesh/xml_data_elements.html.

[31] D. Zagyva, "Project Laboratory Report," Budapest University of Technology and Economics, 2016.

[32] T. Y. K. S. C. T. a. H. T. Y. Beh, "Building classification models from imbalanced fraud detection data".

[33] B. W. R. K. A. R. H. A. A. F. S. K. Z. &. A. N. N. Yap, "An application of oversampling, undersampling, bagging and boosting in handling imbalanced datasets," *Proceedings of the First International Conference on Advanced Data and Information Engineering (DaEng-2013),* pp. 13-22, 2014.

[34] B. S. B. V. V. G. Csaba Kerepesi, "How to Direct the Edges of the Connectomes: Dynamics of the Consensus Connectomes and the Development of the Connections in the Human Brain," *Plos One,* vol. 11.6, 2016.