

RESEARCH ARTICLE

Comparing the Hierarchy of Keywords in On-Line News Portals

Gergely Tibély^{1*}, David Sousa-Rodrigues², Péter Pollner³, Gergely Palla³

1 Department of Biological Physics, Eötvös University, Budapest, Hungary, **2** The Design Group, Faculty of Maths, Computing and Technology, The Open University, Walton Hall, Milton Keynes, United Kingdom, **3** MTA-ELTE Statistical and Biological Physics Research Group, Hungarian Academy of Sciences, Budapest, Hungary

* tibelyg@hal.elte.hu



OPEN ACCESS

Citation: Tibély G, Sousa-Rodrigues D, Pollner P, Palla G (2016) Comparing the Hierarchy of Keywords in On-Line News Portals. PLoS ONE 11 (11): e0165728. doi:10.1371/journal.pone.0165728

Editor: Peter Csermely, Semmelweis University, HUNGARY

Received: August 16, 2016

Accepted: October 17, 2016

Published: November 1, 2016

Copyright: © 2016 Tibély et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: Data is owned by third parties Spiegel Online (www.spiegel.de), The Guardian (www.theguardian.com), The New York Times (www.nytimes.com), The Australian (www.theaustralian.com.au). David Sousa-Rodrigues (david@davidrodrigues.org) is happy to provide the data to all interested researchers upon request, if the owners of the data approve the request. Any interested researcher can obtain the data underlying our study's findings upon request in the same way the authors obtained the data.

Funding: The research was supported by the Hungarian National Science Fund/Országos

Abstract

Hierarchical organization is prevalent in networks representing a wide range of systems in nature and society. An important example is given by the tag hierarchies extracted from large on-line data repositories such as scientific publication archives, file sharing portals, blogs, on-line news portals, etc. The tagging of the stored objects with informative keywords in such repositories has become very common, and in most cases the tags on a given item are free words chosen by the authors independently. Therefore, the relations among keywords appearing in an on-line data repository are unknown in general. However, in most cases the topics and concepts described by these keywords are forming a latent hierarchy, with the more general topics and categories at the top, and more specialized ones at the bottom. There are several algorithms available for deducing this hierarchy from the statistical features of the keywords. In the present work we apply a recent, co-occurrence-based tag hierarchy extraction method to sets of keywords obtained from four different on-line news portals. The resulting hierarchies show substantial differences not just in the topics rendered as important (being at the top of the hierarchy) or of less interest (categorized low in the hierarchy), but also in the underlying network structure. This reveals discrepancies between the plausible keyword association frameworks in the studied news portals.

Introduction

Network science has become ubiquitous in describing and modeling a wide range of phenomena in nature and society, applied in studies of systems ranging from the interactions within cells through transportation systems, the Internet and other technological networks to economic networks, collaboration networks and the society [1, 2]. One of the important topics in this area is related to hierarchical organization of networks [3–12], which is mostly motivated by the fact that signs of hierarchy were recorded in numerous real networks. Examples include various animal flocks [13–16], social interactions [17–19], urban planning [20, 21], ecological systems [22, 23] and evolution [24, 25].

Tudományos Kutatási Alapprogramok, grant no. OTKA K105447 (<http://otka.hu/>). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

A hierarchy is usually depicted as a directed acyclic graph (DAG), where the nodes are layered in different levels, with links pointing from nodes in higher levels towards nodes in lower levels. However, we can distinguish between hierarchies of different types. In a flow hierarchy, the links can represent the chain of command, the flow of information, etc., and in general, they indicate that nodes found in lower levels are influenced by their in-neighbors positioned in higher levels. On the other hand, in a nested hierarchy (also called inclusion hierarchy or containment hierarchy) the links represent an “is a sub-category of” type of relation, as nested hierarchies are obtained by recursively aggregating items into larger and larger groups, resulting in a structure where higher-level groups consist of smaller and more specific components [26]. Prominent examples of nested hierarchy are given by library classification systems and biological classification.

A very interesting problem related to nested hierarchies and classification is given by the automated extraction of nested hierarchies from folksonomies and collaborative tagging systems [27–31]. The association of tags to various on-line contents have become widespread, as various tags may indicate the topic of news-portal feeds and blog post, the genre of films or music records on file sharing portals, or the kind of goods offered in Web stores. These tags usually serve as keywords, providing a rough description of the given entity, helping the users in a fast decision whether the given article, film, etc. is of interest or not. Since the tags appearing in these on-line platforms are usually free words chosen by the author or owner of the given object, they are almost never organized into a pre-defined hierarchy of categories and sub-categories [32–35]. Furthermore, in many tagging systems like Flickr, CiteUlike or Delicious the tagging process is collaborative, as in principle an unlimited number of users can tag photos, Web pages, etc., with free words [36–38]. In order to highlight this collaborative nature, the arising set of free tags and associated objects are often referred to as folksonomies. Since the tagging actions involve user-tag-object triplets, a natural representations of these systems is given by hypergraphs [37, 39–42], where the hyperedges connect more than two nodes together.

In almost all tagging systems the tagged items are allowed to have multiple tags simultaneously, and based on the weighted co-occurrence network between the tags it is possible to extract a nested hierarchy between the tags according to several different algorithms [27–31]. This is usually motivated by the assumption that the way users think about objects or concepts presumably has some built in hierarchy, e.g., “pigeon” is usually considered as a special case of “bird”. Revealing the hidden hierarchy between tags in a folksonomy or in a tagging system in general can significantly help broadening or narrowing the scope of search in the system, give recommendation about yet unvisited objects to the user, or help the categorization of newly appearing objects [42, 43].

Along the above lines, here we apply an improved version of a recent tag hierarchy extraction method [31] to co-occurrence networks between keywords associated to on-line articles, collected from the portals of Spiegel Online, The Guardian, The New York Times and The Australian. The obtained hierarchies show very interesting differences, e.g., the topics rendered as important (being at the top of the hierarchy) in one journal may turn out to be of less interest (categorized lower) in the hierarchy of another journal. In addition, some of the network characteristics of the obtained hierarchies are also quite different, indicating plausible discrepancies in the keyword association frameworks applied in the studied news portals.

Materials and Methods

Hierarchy construction

We employ an upgraded version of a recent method [31], which is based on two assumptions:

- tags positioned high in the hierarchy also have high centrality values in the tag-tag coappearance graph,
- parent-child pairs coappear more frequently than expected from pure chance.

According to the first assumption, the algorithm orders the tags by their centrality, then, for each tag (which become child) the parent candidates are collected. All tags with higher centrality are parent candidates of the child tag. Candidate parents are assigned a score, indicating the probability of the observed number of co-occurrences according to a random null-model. Using the second assumption, the final parent is the candidate with the highest score sum, where the sum runs over all descendants of the child tag. Note, that the algorithm builds up the hierarchy bottom up, starting from the leaves with lowest centrality. The full detailed description of the currently used version of the algorithm involving a couple of improvements is given in [S1 File](#).

Similarity of hierarchies

Hierarchies are frequently represented by Directed Acyclic Graphs (DAGs), in which directed cycles are forbidden. However, children are allowed to have more than one parent in general. For simplicity, we have restricted the number of parents to one in the present analysis. A natural idea for comparing two DAGs is to compare the hierarchical relations, i.e., the sets of ancestor-descendant relationships [31, 44–47]. Here we adopt the approach proposed in Ref. [31], defining a similarity measure based on mutual information. We note that mutual information plays a central role also in the comparison method introduced in [48] for the related, but separate problem of comparing hierarchical community structures, (where only the lowest-level nodes in DAG actually exist in the input data-set). The DAG similarity measure we use, called normalized mutual information (NMI), can be formulated as follows [31]

$$I_{\alpha,\beta} = \frac{2 \sum_{x=1}^{N_{\alpha\beta}} |d_{\alpha}(x) \cap d_{\beta}(x)| \cdot \ln \left(\frac{|d_{\alpha}(x) \cap d_{\beta}(x)|(N-1)}{|d_{\alpha}(x)| \cdot |d_{\beta}(x)|} \right)}{\sum_{x=1}^{N_{\alpha}} |d_{\alpha}(x)| \ln \left(\frac{|d_{\alpha}(x)|}{N-1} \right) + \sum_{x=1}^{N_{\beta}} |d_{\beta}(x)| \ln \left(\frac{|d_{\beta}(x)|}{N-1} \right)} \tag{1}$$

where α and β are two DAGs, having N_{α} and N_{β} tags from which $N_{\alpha\beta}$ are common, $N = N_{\alpha} + N_{\beta} - N_{\alpha\beta}$ is the total number of tags, and $d_{\alpha}(x)$ is the set of descendants of x in DAG α . Eq 1 is 0 for independent DAGs and 1 for identical ones. Note that in the strict mathematical sense, Eq 1 is not a mutual information, however, to avoid confusion, we kept the name used in earlier publications.

A further very closely related similarity measure that turned out to be useful in previous studies is given by the linearized mutual information (LMI) [31], based on the fraction of links that have to be rewired in a randomization procedure on α leading to a hierarchy α_{rand} with the same NMI when compared to α as the $I_{\alpha,\beta}$. The formal definition of this measure is given as follows. Let $I(f)$ denote the average NMI obtained for a fraction of f randomly rewired links, $I(f) = \langle I_{\text{original,rand}} \rangle_f$. By projecting the NMI of the empirical case, I_e , to the f axis using this function as

$$f^* = I^{-1}(I_e), \tag{2}$$

we receive the fraction of randomly chosen links to be rewired in the empirical case for obtaining a randomized hierarchy with the same NMI. Based on that we define the linearized mutual information, (LMI) as

$$I_{\text{lin}} = 1 - f^* = 1 - I^{-1}(I_e) \tag{3}$$

This quality measure corresponds to the fraction of unchanged links in a random link rewiring process, resulting in a hierarchy with the same NMI as the empirical value. (The reason for calling it “linearized” is that Eq 3 is actually projecting I_e to the linear $1 - f$ curve). In theory, one might use the minimum ratio of rewired links which gives the same NMI, however, as [31] showed, change in the NMI can strongly depend on the position of the rewired links, therefore, the minimum number can be significantly lower than the average.

Data

We analyze four tagged datasets, obtained from online news portals. They contain tagged news items, covering a more than 2 years long time window, in the same period. The four sources are: Spiegel Online, The Guardian, The New York Times and The Australian.

General observations. There are some observations which hold for all four datasets. For example, very long tags exist, more like headlines (“Muntazer al-Zaidi: the Iraqi shoe thrower”). Some of the tags form frozen cliques in the coappearance network, where each member of such a clique appear only together with the other members of the clique, e.g., “Haiti” and “Haiti Earthquake Disaster 2010”, “Diana” and “Princess of Wales”. Since members of a large clique have large centrality values, such tags will be placed to unwanted high positions by the first step of the hierarchy construction algorithm. Therefore, we have considered such frozen cliques as single tags, which fits better to the assumed usage of tags.

Some concepts are represented by two or more tags, where the same idea is expressed with different, but synonymous words, e.g., “Art” and “Arts”, “The Arab Revolution” and “Arab Spring” or “Japan disaster” and “Japan earthquake and tsunami”. Since the identification of these more complex tags can be done only by a time-dependent context-aware analysis, they were left as observed, unless explicitly stated otherwise. Another problem is posed by the occurrence of very rare tags, that are usually names.

In order to avoid misleading results due to the above observed problems, we have prefiltered the tags by requiring that each tag pair in the coappearance network has to occur on at least r news items. The $r = 1$ case corresponds to skipping the prefiltering. We set r to its optimal value for each dataset by keeping the number of tags as high as possible and minimizing the number of misleading tags described above. Finally we note, that temporarily important topics can produce unexpected co-occurrences (e.g., “Japan” -> “Fukushima Nuclear Catastrophe” -> “Nuclear Power”).

Spiegel Online. The dataset is from April 2011 to January 2013. It contains 4802 news items and 388 tags. For the pre-filtering, minimum 1 common news item for each tag pair (i.e., no filtering) seems to be a good trade-off between noise reduction and info loss. The dataset looks very well organized (e.g., there are only 400 tags, general tags are used consistently, and there are only a few duplicated tags, long tags or frozen cliques).

The Guardian. The dataset is from November 2009 to January 2013, containing 55835 news items and 6797 tags. Pre-filtering needs minimum 3 news items (removes 2530 tags and 61 news items). Here we found several ad hoc tags (mostly names), that were used only once or a handful of times. We found synonymous tags, e.g., “Middle East and North Africa” and “Middle East”, that will appear as two local roots of two branches in the DAG. These branches correspond to the same topic, thus, divide the related tags between them.

The New York Times. The dataset reaches from November 2010 to January 2013. It contains 35736 news items and 23009 tags. Cliques are a huge problem here. There are 2902 ones, collapsing them removes about 6000 tags. Several cliques appear on numerous objects, therefore, the minimum news item-filtering does not solve the problem automatically. Cliques also

reach very large sizes: there is a 809-tag clique (may contain much more characters than a news item itself); after the minimum news items filtering, the largest one still consists of 44 tags—as follows from the definition of cliques, these tags appear strictly together on each object. For the pre-filtering, minimum 5 news items were required, leaving finally 2981 tags (out of 23009). News items were much less affected, 31184 out of 35736 remained.

The Australian. Data is from December 2009 to January 2013. It contains 31501 news items and 79054 tags—thus, there are much more tags than news items. Cliques are present, but have only 1–2 objects, so it is not a serious problem, the pre-filtering can solve it. Multiple synonyms occur on the same object very often—e.g., “Economist_Paul_Samuelson” “Paul_A._Samuelson” “Paul_Samuelson”. Another example is the set of synonyms for Barack Obama, which are: “Barack_Obama”, “BARACK_Obama”, “Obama”, “PRESIDENT_Barack_Obama”, “President_Barack_Obama”, “President_Obama”, “US_PRESIDENT_Barack_Obama”, “US_President_Barack”, “US_President_Barack_Obama”, “barack_obama”. Pre-filtering with minimum 5 news items leaves 1673 tags out of 79504. The news items are reduced from 31501 to 10550. The tags have relatively few objects, and not only due to the large number of very infrequent tags, e.g., even the prime minister has only 900 objects. Although there are very general tags like “community”, “committee” or “claim”, most of tags are very specific, almost tailored for one object, e.g., “rebels_storm_Gaddafi_compound”.

Results

We analyzed the tag hierarchies obtained from an improved version of “algorithm B” published in Ref. [31]; a brief description of the idea of the method can be found in Materials and Methods, the full details of the used algorithm are given in the [S1 File](#). In Analysis of the individual tag hierarchies first we summarize the most important properties of the individual hierarchies corresponding to the different news portals, which is followed by the pairwise comparisons in Pairwise comparisons. Finally, in Statistical properties of the overall hierarchy structures we examine the overall quality of the hierarchies from different aspects.

Analysis of the individual tag hierarchies

Spiegel Online. The constructed DAG consists of 1 connected component. Most of the tags are under 3 branches: “World”, “Europe”, “Germany”. A visualization of the DAG is shown on [Fig 1](#). The Spiegel DAG seems to be somewhat concerned with immigrants and integration, they have a branch containing 3.9% of the tags, similarly to Australian’s 4.4%, and in contrast to 0.1% and 0.7% of Guardian and NYT (note that the latest data come from January 2013, well before the beginning of the recent migrant crisis).

The Guardian. The overall structure of the DAG is quite well organized, the top 2–3 levels are very impressive. The DAG consists of four similarly-sized connected components: “UK news”, “Worldnews”, “Culture”, “Sport”, although the tags “Worldnews” and “UK news” are in isolated components, they are not completely mutually exclusive, e.g., both of them appear on the news items of “Defence policy”. Note that while the components’ top tags correspond well to the menu items on the journal’s website, they are placed totally automatically by the DAG construction algorithm. Visualization is omitted due to the relatively large size of the DAG, however, a smaller sample is shown on [Fig 2](#).

The New York Times. Here we found numerous duplicated branches in the constructed DAG (e.g., for research, television, education, medicine, defense and military forces). This indicates that for these topics, two distinct sets of tags were used in parallel. The DAG is much less organized than that of the Spiegel and of the Guardian. There are 31 isolated components,

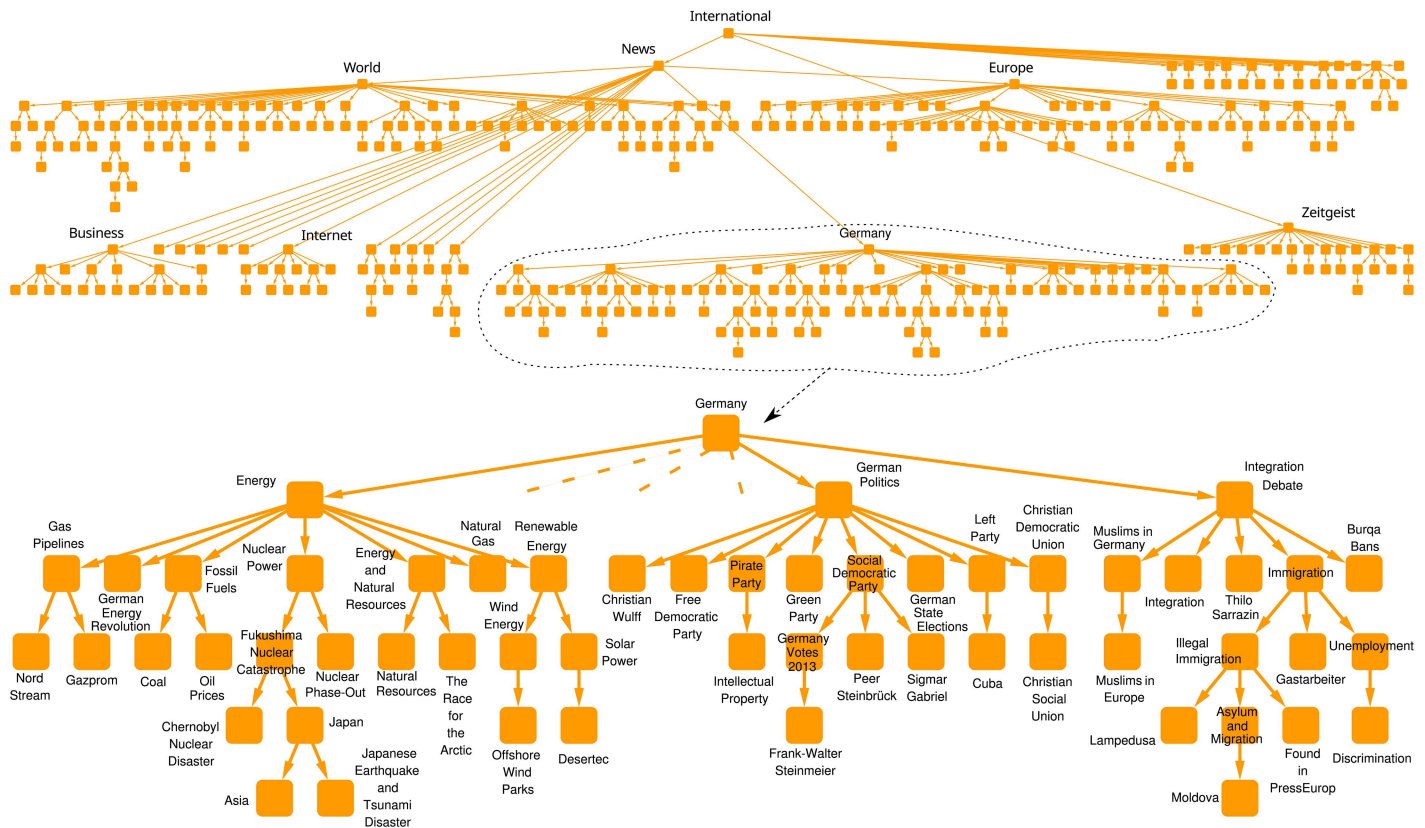


Fig 1. Overview of the Spiegel DAG (top), and one part enlarged (bottom). The DAG is broken into two lines in the top figure to fit the whole graph in the available width. On the bottom figure, dashed links indicate descendants of the tag “Germany” which are not shown.

doi:10.1371/journal.pone.0165728.g001

most of them correspond to one theme (e.g. “Baseball”). The sizes of the components varies from 898 to 2, and there is a continuous range of them from the 2nd largest one (274 tags) down. There are no very general categories. Although a number of large related components exists (under the tags “Basketball”, “Baseball”, “Football”), these components are not collected under a general “Sport” tag. It seems as if there were no demand for using general tags. Note that there is a tag called “sports”, however, it appears only on 5 news items, and it is negligible. A technical consequence is that the DAG construction algorithm does not always select the most general tags as roots, because they lack the important connections to other components. Instead, one of the more specific tags can be selected for a central position, for example, “Middle East and North Africa Unrest (2010–)” for foreign affairs, or “European Sovereign Debt Crisis (2010–)” for Europe-related tags. In other words, the centrality no longer correlates only with the generality for the top tags. Some lower-level branches end up at unexpected places, e.g., “Environment” under “Iran”. Superfluous levels appears, for example, “International Relations” under “United States International Relations”.

The Australian. The DAG looks disorganized overall. There are about 1900 components for the 79504 tags without the pre-filtering, and about 300 components for the min. 5 news items-filtered 1673. There are no macroscopic components, the largest one’s size is just 3480 (out of 79504 tags) and 165 (out of 1673 tags), which is less than 10% of the total nodes. Even the existing components look more like just bunches of more or less associated tags than small hierarchical structures.

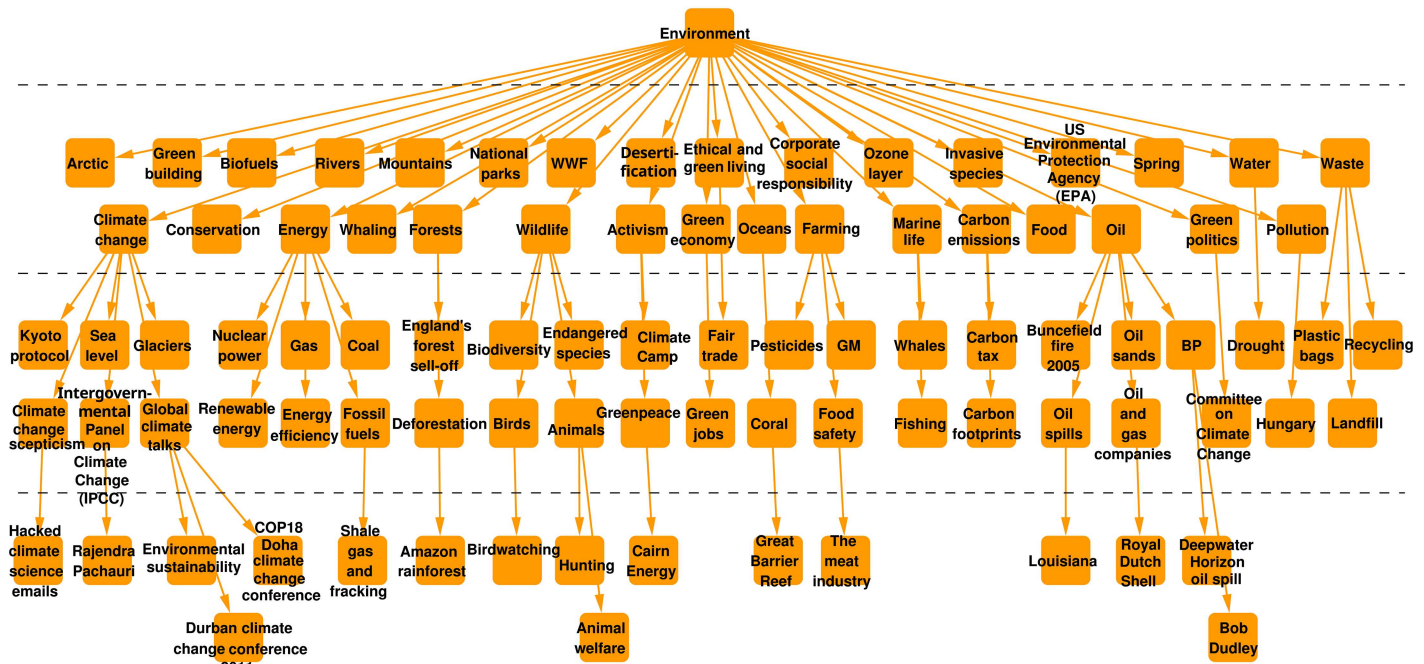


Fig 2. Part of the Guardian's "Environment" branch, in the component "world news". Hierarchical levels are separated by dashed lines.

doi:10.1371/journal.pone.0165728.g002

In general, the top of the constructed DAGs are much better than the bottom. This is no surprise—there is much more information for the construction algorithm at the top of the DAG.

Pairwise comparisons

We carried out a pairwise comparison between the journals from the point of view of their content organization. Since the audience and the interests of the journals are different, the list of tags appearing on the articles was unique for each news portal. Therefore, before actually comparing the tag hierarchies, first we needed to create a common tag set for each pair of journals. In a number of cases, finding the corresponding tag pairs went beyond a simple string matching and was based on semantic matching, e.g., "Fossil fuels" (Guardian) was matched with "Oil (Petroleum) and Gasoline" (NYT). Matching was done manually, which may limit the analysis of datasets larger than the ones presented here. The size of the reduced common tag sets were 252 (Spiegel-Guardian), 217 (Spiegel-NYT), 985 (Guardian-NYT), 93 (Australian-Spiegel), 278 (Australian-Guardian), 274 (Australian-NYT).

The reduced hierarchies were obtained by keeping only the common tags in the original DAGs and erasing the rest of the tags. In most cases this resulted in deletion of leafs, sub-branches, or lower parts of sub-branches from the original hierarchies. However, a small number of times this procedure erased a tag higher in a given branch while keeping other tags lower in the same branch, therefore, distorting the original DAG structure in a radical way. To ensure as much similarity to the original hierarchies as possible, under these circumstances the ancestors standing higher in the branch were also kept, despite that they were not part of the common tag set, (see [S1 File](#) for more details). The reduced DAGs can be found in [S1 File](#).

For each pair of journals we have computed the linearized information similarity measure described in Similarity of hierarchies between the reduced DAGs, the obtained values are shown in [Fig 3](#). According to the results Spiegel and Guardian provide the largest similarity measure, which is also supported by a number of identical or almost identical sub-branches

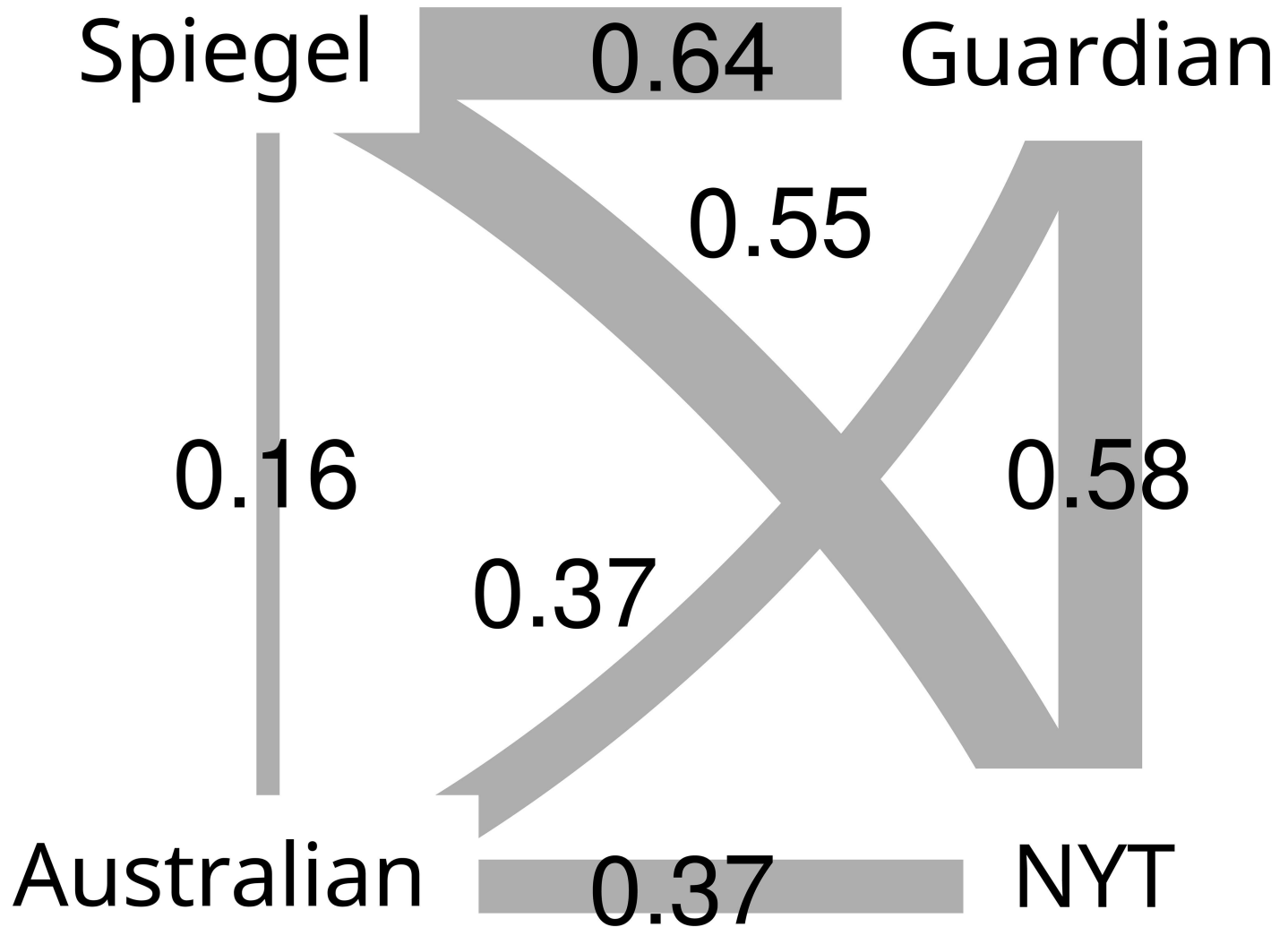


Fig 3. Similarities between the news portals' DAGs, according to the mutual information-based linearized information similarity measure described in Similarity of hierarchies.

doi:10.1371/journal.pone.0165728.g003

between the two DAGs, as shown in Fig 4. Here the background coloring of the sub-branches indicate the similarity to the corresponding (most similar) sub-branch in the other DAG.

The Spiegel, the Guardian and the New York Times have an overall similar structure, as Fig 3 shows, opposed to the Australian, which is dissimilar to all of them. Still, there are some differences between the first three journals. The Guardian, compared to the Spiegel, has a level of intermediately-sized branches, e.g., “law” or “society” in “UK news”. This level is missing from the DAG of Spiegel. Their global DAG structures are shown in Fig 4.

Meanwhile, the New York Times has interestingly no “World” tag, and foreign countries are separated into 4 different branches, in 3 components (see S1 File for more details). Although the linearized information similarity between the Guardian and the New York Times is somewhat lower, they also have a few quite similar branches; a prominent example is shown in Fig 5.

Statistical properties of the overall hierarchy structures

According to the results presented in the previous sections the tag hierarchies obtained for the studied journals show strong differences. Here we examine to what extent does their overall

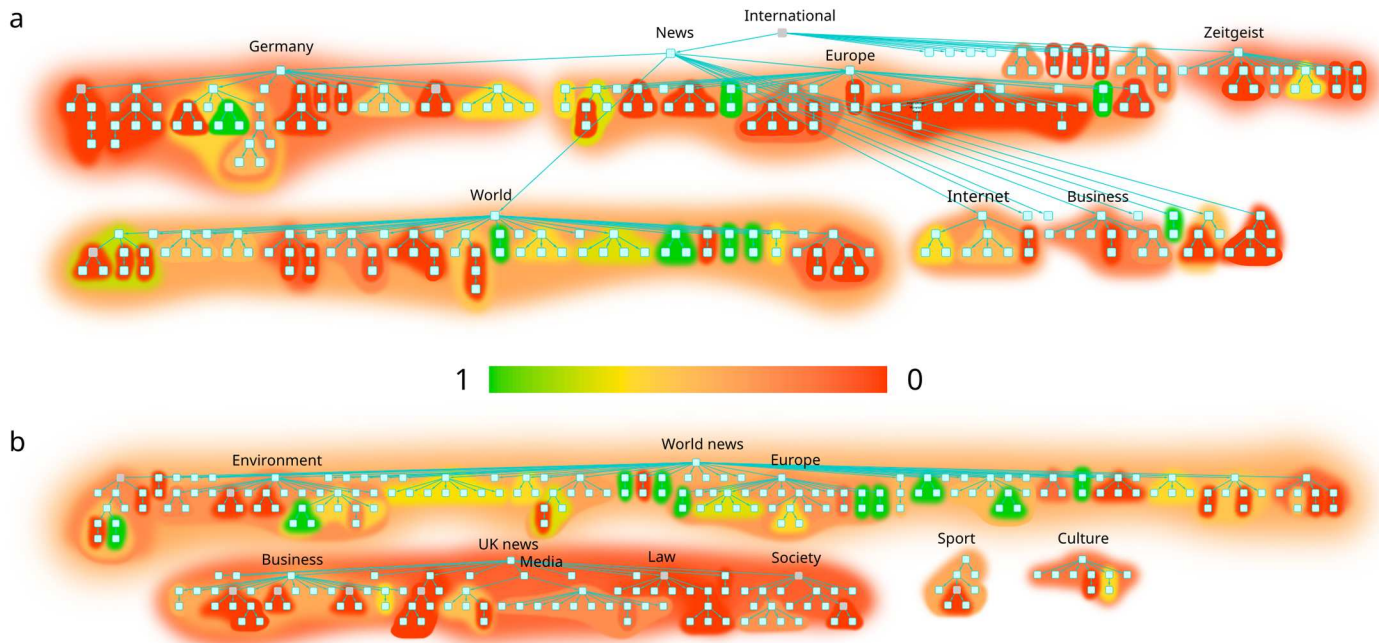


Fig 4. The Spiegel (top) and the Guardian's (bottom) reduced DAG structures, providing the largest overall similarity in our analysis. For clarity, Spiegel's DAG is broken into two lines. Background colors show the result obtained by applying the similarity measure given in Eq 1 to the given branch and the most similar branch from the other hierarchy. Note that sub-branches on all hierarchical levels have their own color.

doi:10.1371/journal.pone.0165728.g004

structure follow a few simple intuitive requirements that can be formulated for a well organized tag hierarchy.

Correlations with Google News. One of the basic properties of a well organized hierarchy is that frequent, more general tags are expected to be higher compared to rare, specific tags. In order to examine the obtained hierarchies from this perspective we compared the centrality score of the tags in the tag co-occurrence network (determining their position in the hierarchy) with their number of hits provided by Google News. For each pair of tags with a significant number of co-occurrence we checked whether the difference between their centrality score and the difference between their number of hits in Google News have the same or the opposite sign. If the signs of the differences match for the majority of the tag pairs, then we can assume that the structure of the hierarchy is consistent with word frequencies of English news texts around the world. Although the magnitudes of the differences are not compared, large relative deviations can also mean that the signs change in other relations of the affected tags, thus, tracking the signs can reveal large relative deviations in certain cases.

In Table 1 we show the relative frequency of the cases, where the differences have the opposite sign, calculated for tag pairs co-appearing in statistically significant numbers. If tags are assigned to articles absolutely at random, the result would correspond to a 0.5 inversion rate, i.e., half of the coappearing tag pairs would have similar centrality and frequency ordering. According to Table 1, the Spiegel and the Guardian data sets provide the best correspondence between tag frequency and centrality, with only a few percent difference in their score. They are followed by the New York Times, and finally, the Australian has a score close to the random case. Although the Google News data may be somewhat different from a fictitious collection word usage of all English speaking journalist, the results in Table 1 show a quite clear-cut picture, which also corresponds well to the results of other comparisons.

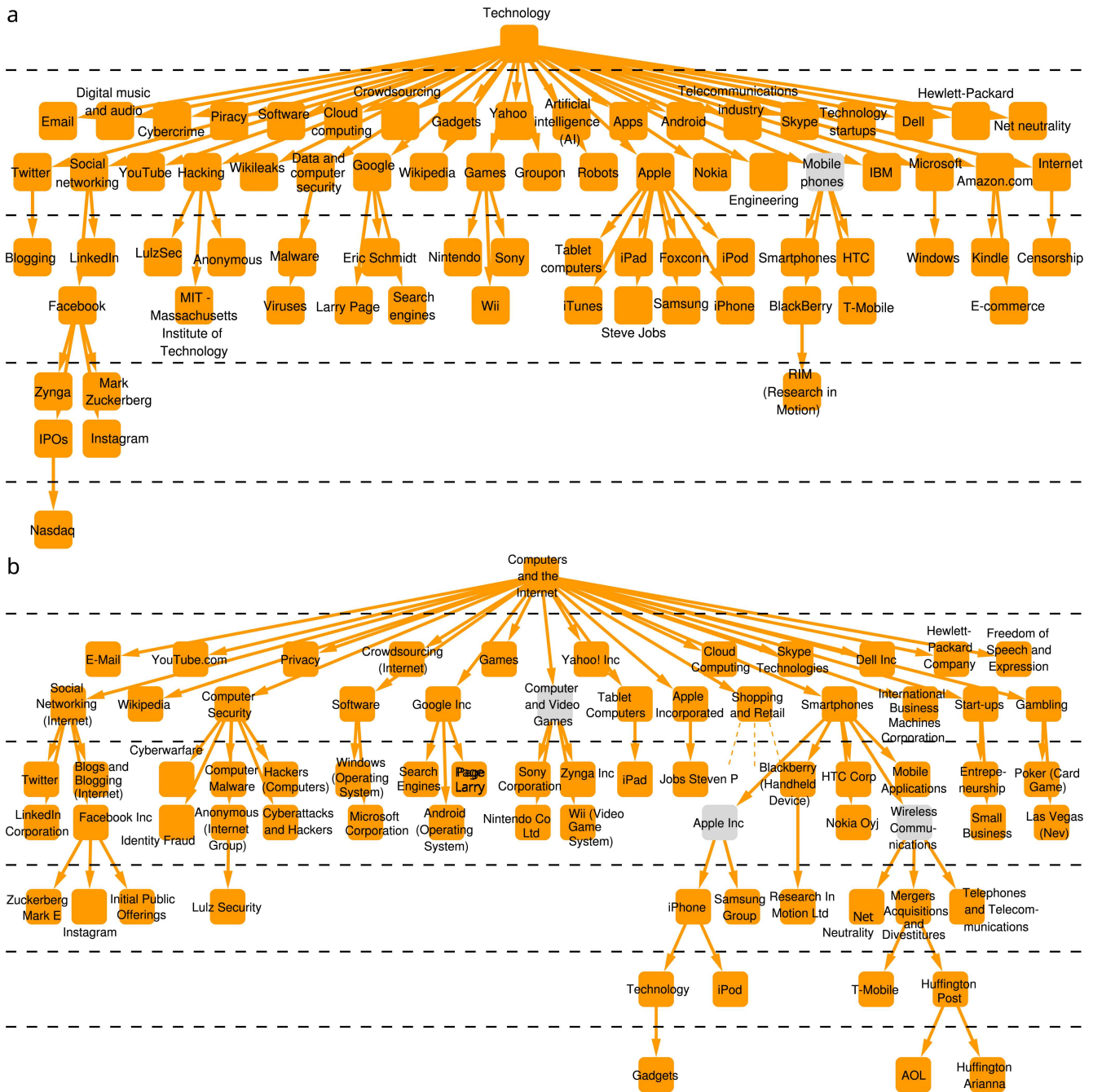


Fig 5. Guardian’s “Technology” branch (top) and New York Times’ “Computers and the Internet” component (bottom). Hierarchical levels are separated by dashed lines. Grey tags do not appear in both DAGs, however, they connect branches containing common tags.

doi:10.1371/journal.pone.0165728.g005

Geometrical properties of the hierarchies. In this section we focus on the geometrical properties of the tag hierarchies from the perspective of whether their structure is helping navigability and search. First we examine the fragmentation of the DAGs, which we can quantify by first introducing the average size of the component of a randomly chosen tag

Table 1. Ratios of inversions between centralities and real-world occurrence frequencies, calculated for tag pairs coappearing in statistically significant numbers. Totally random case corresponds to 0.5.

dataset	ratio of inversions
Spiegel	0.19
Guardian	0.21
New York Times	0.31
Australian	0.44

doi:10.1371/journal.pone.0165728.t001

given by,

$$\tilde{s} = \frac{\sum_i^{\text{tags}} s_i}{N} \tag{4}$$

where s_i is the size of the component containing tag i and N is the total number of tags. Based on \tilde{s} we can calculate the expected lowest hierarchy level l on which the top node of a branch of size \tilde{s} would appear in a balanced k -ary tree of size N . In such a tree any branch can contain at most half of the tags of its mother branch, thus, we define l as

$$\begin{aligned} l &= \lceil \log_2 N / \tilde{s} \rceil, & \tilde{s} < N \\ l &= 1, & \tilde{s} = N \end{aligned} \tag{5}$$

where $\lceil x \rceil$ denotes the ceiling function of x . The value of l becomes high for strongly fragmented tag hierarchies consisting of many small isolated components, where the navigability of the hierarchy is low. The results for \tilde{s} and l are summarized in Table 2. The tag hierarchy obtained for Spiegel (consisting of a single component) provides the lowest l value, followed by Guardian and New York times. Apparently, the DAG of Australian is showing a very fragmented structure with $l = 6$.

Another important question is whether branch sizes are balanced or not in the hierarchies. A well-balanced hierarchy is expected to have at least 2 but not more than $\mathcal{O}(1)$ comparably sized branches at every nonleaf tag. We define a balancedness measure with a pair of real numbers from $[0, 1) \times [0, 1)$ corresponding to the ratio of “giant branches” and the ratio of “dwarf branches” in order to quantify how a DAG fits to the above criterion. First, we calculate the cumulated size of the branches having a child branch which contains more than 50% of the parent branch’s tags. Second, we calculate the cumulated size of the child branches which are smaller than 10% of their parent branches. The higher threshold is motivated by the fact that a child branch above 50% is larger than all the other child branches combined. The motivation for the lower threshold is that below 10%, for equal-sized child branches, the number of child branches exceeds $\mathcal{O}(1)$. Other numerical threshold values might also be applied, however, for demonstrating significant phenomena the precise value of the thresholds should not be important. We normalize the sums by their maximal possible value, thus, our balancedness measure

Table 2. Characteristic level showing the highest level of an idealized hierarchy to which an average connected component corresponds.

dataset	\tilde{s}	N	l
Spiegel	388	388	1
Guardian	1338.7	4263	2
New York Times	384.2	2945	3
Australian	46.2	1487	6

doi:10.1371/journal.pone.0165728.t002

Table 3. Ratios of giant and dwarf branches among all branches, size-weighted.

dataset	R_g	R_d
Spiegel	0.32	0.22
Guardian	0.10	0.42
New York Times	0.42	0.22
Australian	0.26	0.17

doi:10.1371/journal.pone.0165728.t003

is given by

$$(R_g, R_d) = \left(\frac{\sum_g S_g}{\sum_b S_b}, \frac{\sum_d S_d}{\sum_b S_b} \right) \tag{6}$$

where b goes over all branches containing at least 2 tags, S_b is the size of branch b , g goes over branches containing a sub-branch having more than 50% of g 's tags, and S_g is the corresponding branch size, and d goes over sub-branches which are smaller than 10% of their parent branches with S_d being the corresponding branch size. A perfectly balanced hierarchy would have a (0, 0) score and the two extremely unbalanced cases would have (1, 0) for a chain and (0, 1) for a star graph. The results for (R_g, R_d) are given in [Table 3](#).

Spiegel's R_g is dominated by a single contribution. The global root, "International" has a branch containing almost the whole DAG under "News". Most of R_d comes from small branches, although there are a few exceptions. In the Guardian DAG, dwarf sub-branches are common, due to the huge size of the components which dwarf several branches, as well as to nearly star-shaped branches, sometimes containing hundreds of leaf-tags (e.g., "Film", "Music"). For the NYT, contrary to the Guardian, R_g is much larger than R_d . Two important reasons are misplacing a number of branches and letting less general tags getting high centralities. Since the Australian DAG has quite limited structure inside the numerous small components, R_g and R_d are not very informative measures here. However, the tiny components seem to be well balanced.

Further analysis of the DAGs can be found in [S1 File](#).

Discussion

We studied the hierarchy of keywords associated to news articles in four different on-line news portals. The datasets contain various artifacts, such as long and complex keywords, frozen cliques of exclusively coappearing tags, synonyms or very rare and specific tags. Nonetheless, it was possible for the construction method to obtain very reasonable DAGs from the data. The identification of frozen cliques might also be applied by disambiguation techniques, to identify cliques of equivalent semantic meaning, used in the field of Natural Language Processing. The constructed DAGs suggest that the tags appearing in the different news portals are organized to different degrees. Our analysis revealed that Guardian has an extra intermediate level of organization at certain locations. A further very interesting result is that the number of connected components in the DAGs conveys information about the extent of organization in the data: the Spiegel and Guardian have $\mathcal{O}(1)$ components and are quite organized, the New York Times has a few dozen components and breaks the world into independent pieces, and the Australian has $\mathcal{O}(100)$ components which are barely informative at all.

A similar picture was emerging from the comparison between the frequencies of tags in Google News and their centrality score in the tag-tag co-appearance graphs. The correlation was quite strong in case of the Spiegel and the Guardian, medium for the New York Times, and almost equivalent to the totally random case for the Australian. A more detailed

characterization of the DAGs can be obtained by quantifying the extents of too large and too small sub-branches. Although being a geometry-based analysis, it can also identify problems with tag functions, like a non-comprehensive set of intermediate-level branches in the Guardian, or misplaced branches in the New York Times.

In summary, the following picture is arising from the different analyses we carried out: the Spiegel and Guardian datasets are quite well-organized, the New York Times is significantly less but still has relevant hierarchical structure, and the Australian is close to being random, from a hierarchical point of view. The consistency of the results is encouraging, and suggests that the measures used are useful in the quantification and comparison of datasets from the aspect of hierarchical organization.

Supporting Information

S1 File. S1 File provides more details on the applied algorithm and on the pairwise comparisons of the DAGs.

(PDF)

Acknowledgments

The authors thank the Spiegel Online, The Guardian, The New York Times and The Australian for allowing access to their data.

Author Contributions

Conceptualization: GT PP GP.

Data curation: DSR.

Formal analysis: GT GP.

Funding acquisition: GP PP.

Investigation: GT PP GP.

Methodology: GT PP GP.

Project administration: GP PP.

Resources: DSR.

Software: GT.

Supervision: GP.

Visualization: GT GP.

Writing – original draft: GT GP.

Writing – review & editing: GT GP PP.

References

1. Barabási AL. Network Science. Cambridge Univ. Press; 2016.
2. Dorogovtsev SN, Mendes JFF. Evolution of Networks: From Biological Nets to the Internet and WWW. Oxford Univ. Press; 2014.
3. Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabási AL. Hierarchical Organization of Modularity in Metabolic Networks. Science. 2002; 297:1551–1555. doi: [10.1126/science.1073374](https://doi.org/10.1126/science.1073374) PMID: [12202830](https://pubmed.ncbi.nlm.nih.gov/12202830/)

4. Trusina A, Maslov S, Minnhagen P, Sneppen K. Hierarchy measures in complex networks. *Phys Rev Lett*. 2004; 92:178702. doi: [10.1103/PhysRevLett.92.178702](https://doi.org/10.1103/PhysRevLett.92.178702) PMID: [15169201](https://pubmed.ncbi.nlm.nih.gov/15169201/)
5. Clauset A, Moore C, Newman MEJ. Hierarchical structure and the prediction of missing links in networks. *Nature*. 2008; 453:98–101. doi: [10.1038/nature06830](https://doi.org/10.1038/nature06830) PMID: [18451861](https://pubmed.ncbi.nlm.nih.gov/18451861/)
6. Pumain D. Hierarchy in Natural and Social Sciences. vol. 3 of Methodos Series. Dordrecht, The Netherlands: Springer; 2006.
7. Corominas-Murtra B, Rodríguez-Caso C, Goñi J, Solé R. Measuring the hierarchy of feedforward networks. *Chaos*. 2011; 21:016108. doi: [10.1063/1.3562548](https://doi.org/10.1063/1.3562548) PMID: [21456850](https://pubmed.ncbi.nlm.nih.gov/21456850/)
8. Mones E, Vicsek L, Vicsek T. Hierarchy Measure for Complex Networks. *PLoS ONE*. 2012; 7:e33799. doi: [10.1371/journal.pone.0033799](https://doi.org/10.1371/journal.pone.0033799) PMID: [22470477](https://pubmed.ncbi.nlm.nih.gov/22470477/)
9. Corominas-Murtra B, Goñi J, Solé RV, Rodríguez-Caso C. On the origins of hierarchy in complex networks. *Proc Natl Acad Sci USA*. 2013; 110:13316–13321. doi: [10.1073/pnas.1300832110](https://doi.org/10.1073/pnas.1300832110) PMID: [23898177](https://pubmed.ncbi.nlm.nih.gov/23898177/)
10. Czégel D, Palla G. Random walk hierarchy measure: What is more hierarchical, a chain, a tree or a star? *Sci Rep*. 2015; 5:17994. doi: [10.1038/srep17994](https://doi.org/10.1038/srep17994) PMID: [26657012](https://pubmed.ncbi.nlm.nih.gov/26657012/)
11. Kaiser M, Hilgetag CC, Kötter R. Hierarchy and dynamics of neural networks. *Front Neuroinform*. 2010; 4:112. doi: [10.3389/fninf.2010.00112](https://doi.org/10.3389/fninf.2010.00112) PMID: [20844605](https://pubmed.ncbi.nlm.nih.gov/20844605/)
12. Ma HW, Buer J, Zeng AP. Hierarchical structure and modules in the Escherichia coli transcriptional regulatory network revealed by a new top-down approach. *BMC Bioinform*. 2004; 5:199. doi: [10.1186/1471-2105-5-199](https://doi.org/10.1186/1471-2105-5-199) PMID: [15603590](https://pubmed.ncbi.nlm.nih.gov/15603590/)
13. Goessmann C, Hemelrijk C, Huber R. The formation and maintenance of crayfish hierarchies: behavioral and self-structuring properties. *Behav Ecol Sociobiol*. 2000; 48:418–428. doi: [10.1007/s002650000222](https://doi.org/10.1007/s002650000222)
14. Nagy M, Ákos Z, Biro D, Vicsek T. Hierarchical group dynamics in pigeon flocks. *Nature*. 2010; 464:890–893. doi: [10.1038/nature08891](https://doi.org/10.1038/nature08891) PMID: [20376149](https://pubmed.ncbi.nlm.nih.gov/20376149/)
15. Nagy M, Vásárhelyi G, Pettit B, Roberts-Mariani I, Vicsek T, Biro D. Context-dependent hierarchies in pigeons. *Proc Natl Acad Sci USA*. 2013; 110:13049–13054. doi: [10.1073/pnas.1305552110](https://doi.org/10.1073/pnas.1305552110) PMID: [23878247](https://pubmed.ncbi.nlm.nih.gov/23878247/)
16. Fushing H, McAssey MP, Beisner B, McCowan B. Ranking network of captive rhesus macaque society: A sophisticated corporative kingdom. *PLoS ONE*. 2011; 6:e17817. doi: [10.1371/journal.pone.0017817](https://doi.org/10.1371/journal.pone.0017817) PMID: [21423627](https://pubmed.ncbi.nlm.nih.gov/21423627/)
17. Guimerà R, Danon L, Díaz-Guilera A, Giralt F, Arenas A. Self-similar community structure in a network of human interactions. *Phys Rev E*. 2003; 68:065103. doi: [10.1103/PhysRevE.68.065103](https://doi.org/10.1103/PhysRevE.68.065103) PMID: [14754250](https://pubmed.ncbi.nlm.nih.gov/14754250/)
18. Pollner P, Palla G, Vicsek T. Preferential attachment of communities: The same principle, but a higher level. *Europhys Lett*. 2006; 73:478–484. doi: [10.1209/epl/i2005-10414-6](https://doi.org/10.1209/epl/i2005-10414-6)
19. Valverde S, Solé RV. Self-organization versus hierarchy in open-source social networks. *Phys Rev E*. 2007; 76:046118. doi: [10.1103/PhysRevE.76.046118](https://doi.org/10.1103/PhysRevE.76.046118) PMID: [17995071](https://pubmed.ncbi.nlm.nih.gov/17995071/)
20. Krugman PR. Confronting the mystery of urban hierarchy. *J Jpn Int Econ*. 1996; 10:399–418. doi: [10.1006/jjie.1996.0023](https://doi.org/10.1006/jjie.1996.0023)
21. Batty M, Longley P. *Fractal Cities: A Geometry of Form and Function*. San Diego: Academic; 1994.
22. Hirata H, Ulanowicz R. Information theoretical analysis of the aggregation and hierarchical structure of ecological networks. *J Theor Biol*. 1985; 116:321–341. doi: [10.1016/S0022-5193\(85\)80271-X](https://doi.org/10.1016/S0022-5193(85)80271-X)
23. Wickens J, Ulanowicz R. On quantifying hierarchical connections in ecology. *J Soc Biol Struct*. 1988; 11:369–378. doi: [10.1016/0140-1750\(88\)90066-8](https://doi.org/10.1016/0140-1750(88)90066-8)
24. Eldredge N. *Unfinished Synthesis: Biological Hierarchies and Modern Evolutionary Thought*. New York: Oxford Univ. Press; 1985.
25. McShea DW. The hierarchical structure of organisms. *Paleobiology*. 2001; 27:405–423. doi: [10.1666/0094-8373\(2001\)027%3C0405:THSOOA%3E2.0.CO;2](https://doi.org/10.1666/0094-8373(2001)027%3C0405:THSOOA%3E2.0.CO;2)
26. Wimberley ET. *Nested ecology. The place of humans in the ecological hierarchy*. Baltimore, MD: John Hopkins University Press; 2009.
27. Heymann P, Garcia-Molina H. *Collaborative Creation of Communal Hierarchical Taxonomies in Social Tagging Systems*. Stanford InfoLab; 2006.
28. Plangprasopchok A, Lerman K, Getoor L. A Probabilistic Approach for Learning Folksonomies from Structured Data. In: *Fourth ACM International Conference on Web Search and Data Mining (WSDM)*; 2011. p. 555–564.

29. Schmitz P. Inducing Ontology from Flickr Tags. In: Proc. of Collaborative Web Tagging Workshop at the 15th Int. Conf. on World Wide Web (WWW); 2006.
30. Damme CV, Hepp M, Siorpaes K. FolksOntology: An Integrated Approach for Turning Folksonomies into Ontologies. *Soc Networks*. 2007; 2:57–70.
31. Tibély G, Pollner P, Vicsek T, Palla G. Extracting Tag Hierarchies. *PLoS ONE*. 2013; 8:e84133. doi: [10.1371/journal.pone.0084133](https://doi.org/10.1371/journal.pone.0084133) PMID: [24391901](https://pubmed.ncbi.nlm.nih.gov/24391901/)
32. Mika P. Ontologies are us: a unified model of social networks and semantics. In: Gil Y, Motta E, Benjamins VR, Musen MA, editors. *The Semantic Web—ISWC 2005*. vol. 3729 of *Lect. Notes Comput. Sci*. Springer; 2005. p. 522–536. doi: [10.1007/11574620_38](https://doi.org/10.1007/11574620_38)
33. Spyns P, Moor AD, Vandenbussche J, Meersman R. From Folksonomies to Ontologies: How the Twain Meet. In: Meersman R, Tari Z, editors. *On The Move To Meaningful Internet Systems 2006: CoopIS, DOA, GADA, and ODBASE*. vol. 4275 of *Lect. Notes Comput. Sci*. Springer; 2006. p. 738–755. doi: [10.1007/11914853_45](https://doi.org/10.1007/11914853_45)
34. Voss J. Tagging, Folksonomy & Co—Renaissance of Manual Indexing?; 2007.
35. Tibély G, Pollner P, Vicsek T, Palla G. Ontologies and tag-statistics. *New J Phys*. 2012; 14:053009. doi: [10.1088/1367-2630/14/5/053009](https://doi.org/10.1088/1367-2630/14/5/053009)
36. Cattuto C, Loreto V, Pietronero L. Semiotic dynamics and collaborative tagging. *Proc Natl Acad Sci USA*. 2007; 104:1461–1464. doi: [10.1073/pnas.0610487104](https://doi.org/10.1073/pnas.0610487104) PMID: [17244704](https://pubmed.ncbi.nlm.nih.gov/17244704/)
37. Lambiotte R, Ausloos M. Collaborative Tagging as a Tripartite Network. In: Alexandrov VN, van Albada GD, Sloot PMA, Dongarra J, editors. *Computational Science—ICCS 2006*. vol. 3993 of *Lect. Notes Comput. Sci*. Springer; 2006. p. 1114–1117. doi: [10.1007/11758532_152](https://doi.org/10.1007/11758532_152)
38. Cattuto C, Barrat A, Baldassarri A, Schehr G, Loreto V. Collective dynamics of social annotation. *Proc Natl Acad Sci USA*. 2009; 106:10511–10515. doi: [10.1073/pnas.0901136106](https://doi.org/10.1073/pnas.0901136106) PMID: [19506244](https://pubmed.ncbi.nlm.nih.gov/19506244/)
39. Ghosal G, Zlatić V, Caldarelli G, Newman MEJ. Random hypergraphs and their applications. *Phys Rev E*. 2009; 79:066118. doi: [10.1103/PhysRevE.79.066118](https://doi.org/10.1103/PhysRevE.79.066118) PMID: [19658575](https://pubmed.ncbi.nlm.nih.gov/19658575/)
40. Zlatić V, Ghosal G, Caldarelli G. Hypergraph topological quantities for tagged social networks. *Phys Rev E*. 2009; 80:036118. PMID: [19905191](https://pubmed.ncbi.nlm.nih.gov/19905191/) doi: [10.1103/PhysRevE.80.036118](https://doi.org/10.1103/PhysRevE.80.036118)
41. Floeck F, Putzke J, Steinfels S, Fischbach K, Schoder D. Imitation and Quality of Tags in Social Bookmarking Systems—Collective Intelligence Leading to Folksonomies. In: Bastiaens TJ, Baumöl U, Krämer BJ, editors. *On Collective Intelligence*. vol. 76 of *Advances in Intelligent and Soft Computing*. Springer; 2011. p. 75–91.
42. Lu L, Medo M, Yeung CH, Zhang YC, Zhang ZK, Zhou T. Recommender systems. *Phys Rep*. 2012; 519:1–49. doi: [10.1016/j.physrep.2012.02.006](https://doi.org/10.1016/j.physrep.2012.02.006)
43. Juszczyszyn K, Kazienko P, Katarzyna M. Personalized Ontology-Based Recommender Systems for Multimedia Objects. In: Håkansson A, Hartung R, Nguyen N, editors. *Agent and Multi-agent Technology for Internet and Enterprise Systems*. vol. 289 of *Studies in Computational Intelligence*. Springer; 2010. p. 275–292. doi: [10.1007/978-3-642-13526-2_13](https://doi.org/10.1007/978-3-642-13526-2_13)
44. Fattore M, Grassi R, Arcagni A. Measuring Structural Dissimilarity Between Finite Partial Orders. In: Brüggemann R, Carlsen L, Wittmann J, editors. *Multi-indicator Systems And Modelling In Partial Order*. Springer; 2013. p. 69–84.
45. Brandenburg FJ, Gleißner A, Hofmeier A. The nearest neighbor Spearman footrule distance for bucket, interval, and partial orders. *J Comb Optim*. 2013; 26:310–332. doi: [10.1007/s10878-012-9467-x](https://doi.org/10.1007/s10878-012-9467-x)
46. Palla G, Tibély G, Mones E, Pollner P, Vicsek T. Hierarchical networks of scientific journals. *Palgrave Communications*. 2015; 1:15016. doi: [10.1057/palcomms.2015.16](https://doi.org/10.1057/palcomms.2015.16)
47. Tibély G, Pollner P, Palla G. Partial order similarity based on mutual information; 2016.
48. Perotti JI, Tessone CJ, Caldarelli G. Hierarchical mutual information for the comparison of hierarchical community structures in complex networks. *Phys Rev E*. 2015; 92:062825. doi: [10.1103/PhysRevE.92.062825](https://doi.org/10.1103/PhysRevE.92.062825) PMID: [26764762](https://pubmed.ncbi.nlm.nih.gov/26764762/)