Neurological Disorders and Publication Abstracts Follow Social Network-Type Node Patterns When Indexed Using Ontology Tree-Based Key Term Search

Anand Kulanthaivel[1,2,*], Josette F. Jones[1], Robert P. Light[2], Katy Börner[2], Chin H. Kong[2]

[1] Indiana University Indianapolis (IUPUI), School of Informatics & Computing (BioHealth Informatics), Indianapolis, USA
[2] Indiana University Bloomington, Cyberinfrastracture for Network Science, School of Informatics & Computing (Information & Library Science), Bloomington, IN, USA

[*]akulanth@indiana.edu, jofjones@iupui.edu, lightr@indiana.edu, katy@indiana.edu, kongch@indiana.edu

Abstract. Disorders of the Central Nervous System (CNS) are worldwide causes of morbidity and mortality. In order to further investigate the nature of the CNS research, we generate from an initial reference a controlled vocabulary and ontological tree structure for this vocabulary, and then apply the vocabulary in an analysis of the past ten years of abstracts (n = 10,468) from a major neuroscience journal. Using naïve search methodology with our terminology tree, we find over 4,500 relationships between abstracts and clinical diagnostic topics. After generating a network graph of these document-topic relationships, we find that this network graph contains characteristics of document-author and other human social networks, including evidence of scale-free and power law-like node distributions. Lastly, we discuss potential health consumer-centered uses for our ontology and search methodology.

Keywords: Ontology, information retrieval, neuroscience, networks, indexing, knowledge gaps, data mining, semantic medicine, translational medicine

1 Introduction

Research in the field of biomedical science associating publications with explicit clinical diagnostic terms is lacking. While central nervous system (CNS) disorders are a major cause of morbidity and mortality worldwide, there have been no studies to date on correlates between clinical and basic neuroscience terminology.

Given a controlled vocabulary (CV) whose members are organized into a tree-structured ontology, it is possible to search for biomedical or clinical meaning in a corpus of abstracts or other publication identifiers[1,2]. If such an analysis is performed, one result may be the return of another ontology (this time, document-to-topic). The properties of such a network, as with any network, may be explored using basic social graph metrics[3].

Degree-based centrality (connectedness) is one measure of the influence of a node. Distributions of node centralities (including node degrees) have been postulated to allow conclusions to be drawn about a network in general given its centrality distributions[4]; Barabasi[5] in particular states that social-like network distributions, such as the power law distribution, are seen in a

variety of situations that extend beyond sociology. Milojevic[6] proposes that modifications of power laws are allowed.

Therefore, a study of neuroscience publications with respect to the clinical topics that is likely to yield useful results as to the *aboutness* of neuroscience articles and also reveal any topic bias(es). In particular, *Brain Research*, one particularly influential neuroscience journal with over 55,000 publications to date[7], provides an exemplar for a publication network of CNS-related topics. In this study, the past ten years (2004-2013) of abstracts written for *Brain Research* are analyzed against clinical terms from the Merck Manual for Professionals (in particular, the sections on neurological[8] and psychiatric[9] disorders). We may therefore be able to decide if, based upon the analysis of the resulting network, if said network possesses properties of scale-free networks, including, but not limited to power-law centrality distributions.

2 Methods & Materials

2.1 Ontology Construction

The ontology and controlled vocabulary used in this study was derived from the Merck Manual for Professionals, particularly the sections on CNS pathologies[8][9]. From this source, we found ninety-six (96) unique disorders. Each disorder super-heading was made into one diagnostic entity. Using the discretion of the authors, discrete and exclusive key words and key terms were mapped to each diagnostic entity. Therefore, a structure of disorder-to-keywords was created for each disorder. One example of a disorder-keywords tree as used here is seen in Figure 1.

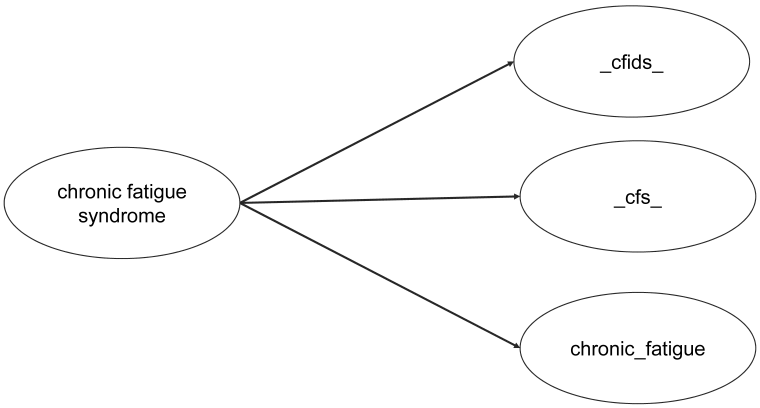

Figure 1. Ontology map visualized (example). Chronic fatigue syndrome is the diagnosis, and the entities it points to are the machine-searchable key terms that this diagnosis maps to. Note that an entire map of our ontology would consist of 96 discrete trees.

2.2 Querying & Basic Information Retrieval

Information retrieval was performed by using the query term *Brain Research[journal]* in PubMed[10]. In order to represent the most recent cohort of documents, the search was filtered to only include articles published from 2004-2013. The result set was downloaded in XML format, and a raw corpus data file created using the Python scripting language. The output generated by the Python script formatted the corpus as *PMID,abstract* where PMID is the PubMed Identifier (PMID) for each article and *abstract* represents the abstract text of the article. Furthermore, in order to enhance machine parsing, all punctuation *within* abstract texts was removed and replaced with the underscore (_) symbol.

A separate file was created in order to contain the ontology trees. The format for each individual disorder tree was *disorder:key_term_1,key_term_2,* with each disorder having one or more key terms. For example, the ontology tree visualized in the above Figure 1 would have been represented in our search word file as *chronic_fatigue_syndrome:_cfs_,_cfids_,myalgic_encephalitis,chronic_fatigue*. Note the underscores surrounding acronyms; these are used to exclude words that might contain these strings as substrings (e.g., the disorder amyotrophic lateral sclerosis [ALS] causing a false positive match in an abstract containing the string f*als*e if *ALS* were a search term used to index the document to that particular disorder).

2.3 Parsing & Knowledge Synthesis

In order to create a graph-like representation of our subject-object construction (and in turn, discover which abstracts were related to which disorders, thus creating knowledge), the PMID/abstract output file was searched against the ontology tree file, and positive matches sent to output in a network tool-readable edge list.

For this purpose, the authors wrote a custom program in Java (Virtual Machine; JVM) using the Eclipse IDE software tool[11]. The search algorithm utilized was naïve, searching explicitly through the corpus file for disorder key words. As output, the algorithm generated an edge list file, with each line being an edge, the left node being the PMID, and the right node being the disorder topic that the matching key term was mapped to in the ontology tree file. For example, an abstract identified by PMID 99999999 and containing key terms for chronic fatigue syndrome, bipolar disorder, and adverse drug reactions would generate the edge list table that follow in Table 1.

Table 1. Sample edge list output for network analysis.

PMID,disorder
99999999,chronic_fatigue_syndrome
99999999,bipolar_disorder
99999999,adverse_drug_reactions

Of remark is that our algorithm was able to avoid parallel edges while constructing an edge list: Should the above document have contained *cfs, cfids,* and *chronic fatigue*, our algorithm will only output the pair *99999999-chronic_fatigue_syndrome* once. This referential integrity was enforced by creating a step where JVM would store the previous keyword term match and refuse to

generate a duplicate edge if the previous key term's parent diagnosis matched any other key term while the algorithm was searching for key terms of that particular diagnosis in that particular abstract.

2.4 Graph Visualization & Metrics

The Sci2 software tool[12] was utilized for initial visualization and metrics computing. Specifically, the DrL layout[13] was used in order to gravitate the node positions for better viewing. Sci2 was then used to compute various centrality measures and the distributions of these measures. This study is limited to the exploration of node degree metrics.

Correlations between degree measures were performed by exporting Sci2-generated data tables into Microsoft Excel[14] and analyzing and plotting the data via the IBM SPSS[15] tool for the disorder degree-rank scatter plot and OpenOffice 3.0[16] for the distribution histograms. Whenever necessary, logarithmic and other curvilinear transformations of data were performed in order to find an ideal fit in the realm of potential curvilinear combinations.

3 Results & Conclusions

3.1 Match rates

Publication-Terminology Match Rate. Recall from our abstract that we searched 10,468 papers (i.e., the result set returned from *Brain Research* as *[journal]* term in PubMed, with the range set to *past 10 years*, and papers that only have available abstracts). We noted that 3,862 papers (34.25%) had abstracts that matched the key terms in our ontology, yielding a corresponding miss rate of 65.75%. However, due to the limited terminology set of the current ontology tree, we chose to use graph analysis for the sub-corpus of publications whose abstracts did match our tree.

One must nonetheless realize that the low match rate, despite our best efforts in engineering the ontology for matching documents, may point to a disconnect (or knowledge gap) between science and medicine. On the other hand, there exists the possibility that many studies are carried out in order to study the normal functioning in the CNS (as opposed to disorder or pathology).

Disorder-Terminology Match Rate. Out of the 96 disorders found in the CNS section of the Merck Manual, 68 of these matched with publications via our ontology tree, yielding a disorder match rate of 70.8%.

3.2 Network Visualization

Network visualization yielded 11 graph components, with the giant component holding 99.2% of all nodes. Therefore, visualization is only focused on the giant component and not the fractured isolates. Figure 2 contains the visualization of the network graph.

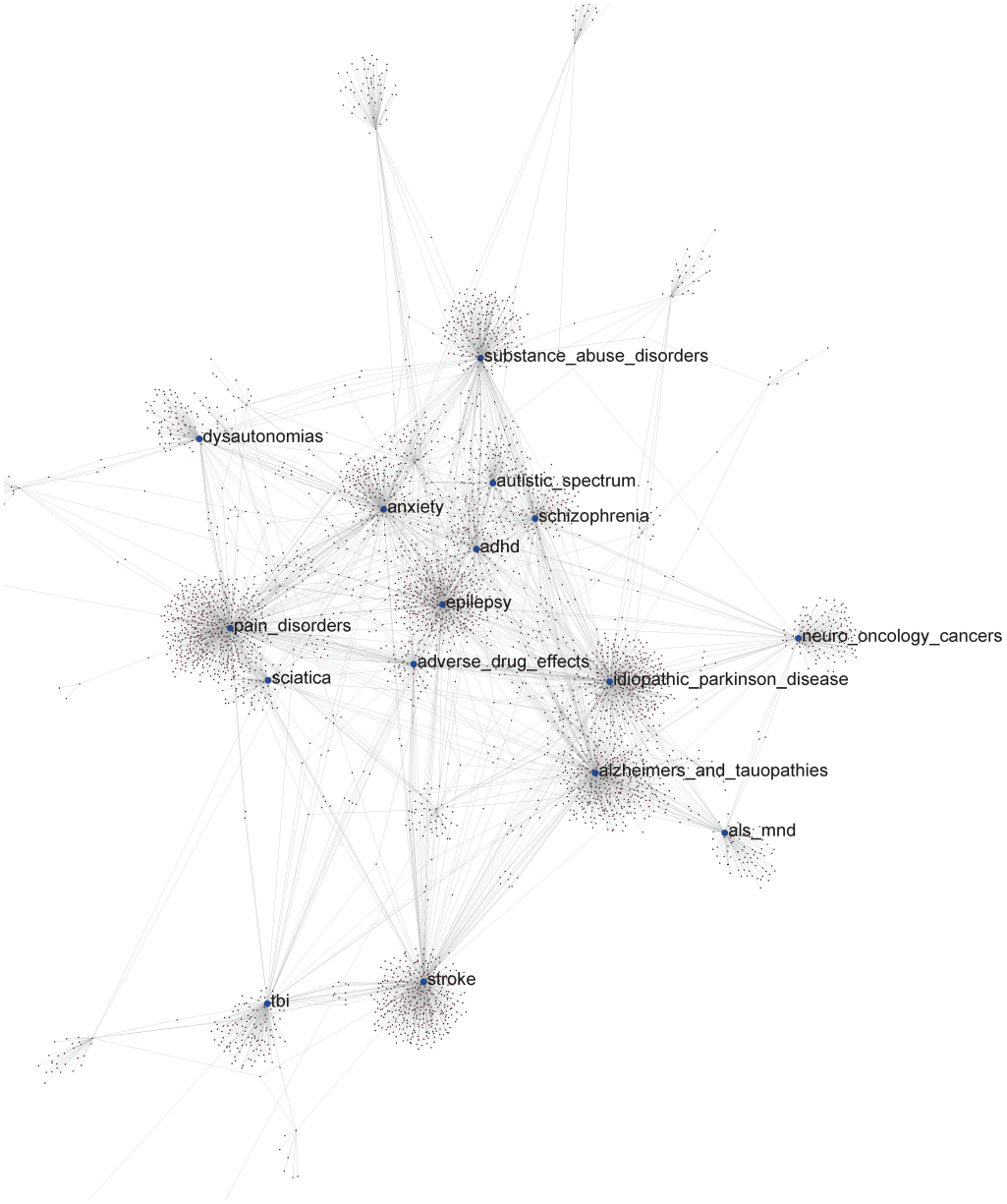

Figure 2. Full graph visualization of network giant component. Some high-profile disorders (Degree Centrality > 100) are highlighted by visible text labels.

3.3 Node Centrality Analysis

Analysis of Disorder Centrality: Degree. The degree (number of publications connected to) of each of the discovered disorders we chose is plotted against the degree rank of these disorders in Figure 3. The data best fit a logarithmic-linear (log-lin) format as per Pearson regression analysis in IBM SPSS[15] (RSQ = 0.985; N = 68). Such a fit suggests a Pareto Type-2 distribution[6]. The most frequently matched disorder (pain disorder) was linked to 587 publications, while several disorders (e.g. bulimia) were linked only to one publication each. As per research conducted by Barabasi[5], scale-free network topology is suggested.

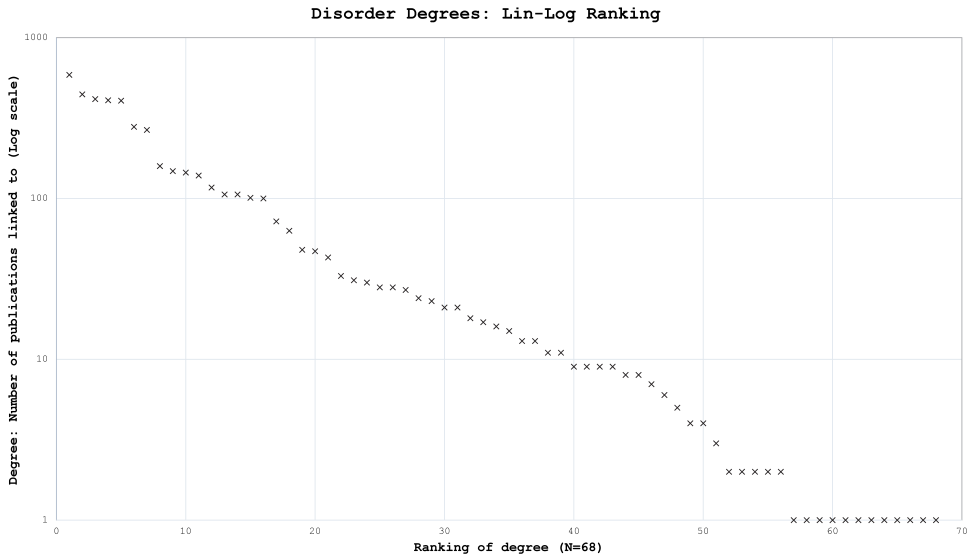

Figure 3. Disorder degree-to-rank correlations. Rank is plotted on the X axis, while disorder degree (connectedness to publications) is plotted on the Y axis. The Y axis is scaled logarithmically as per prescription of Milojevic's publication[6].

Furthermore, when the publication frequency of disorders was binned for histogram analysis, the most parsimonious fit to a typical Z-type distribution appearance was obtained using not linearly-sized bins, but instead, bins sized to powers of *e* (Euler's number; i.e., exponential binning was used). However, it is still of note that while the *e*-power bin histogram distribution appears largely normal, that there is a sharp peak of papers in the first bin. This peak represents disorders connected to between e^0 (i.e., one) and e^1 (i.e., ~2.78) publications. It should be noted, however, that the data ceiling of this bin is technically two (2), as publications are connected to a discrete number of disorders in our model. This relationship is visualized in Figure 4. It is important to note that this visualization only visualizes disorders that matched *any* publication; thus, 28 disorders are excluded as there is no bin for zero (0) to one (1) publications.

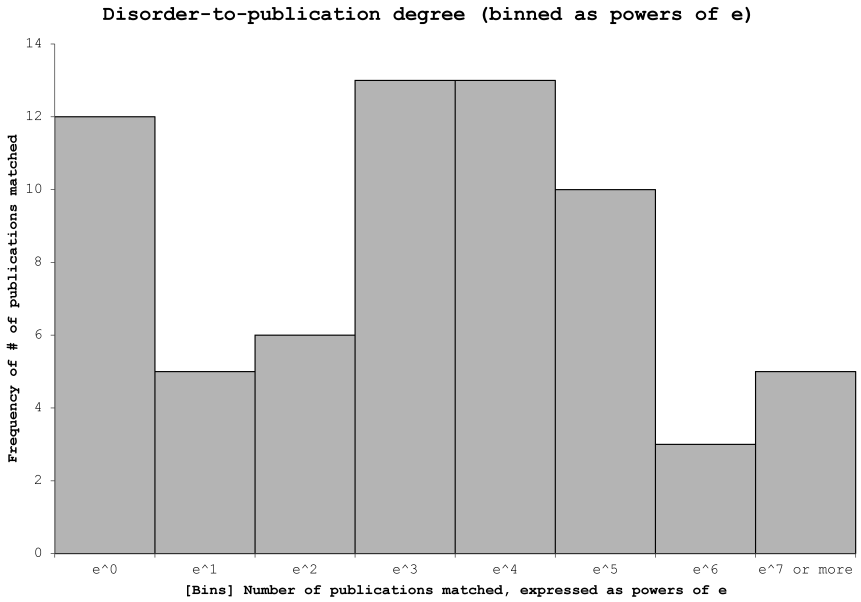

Figure 4. Exponentially-binned histogram (powers of Euler's number) of disorder-to-publication degree distribution, shown on a linear frequency axis; disorders of degree zero cannot be scaled exponentially and are omitted

Analysis of Publication Degrees. Similarly, degree for each publication was analyzed (i.e., how many disorders each publication would connect to). It appears that most publications were connected to only one disorder, while a few yielded matches with several disorders. The highest number of disorders matched for any abstract was six (6), with three (3) abstracts matching six disorders each. A logarithmically-binned histogram shows the logarithmic-linear trend in decreasing frequency of publications with respect to increasing topic degrees (Figure 5).

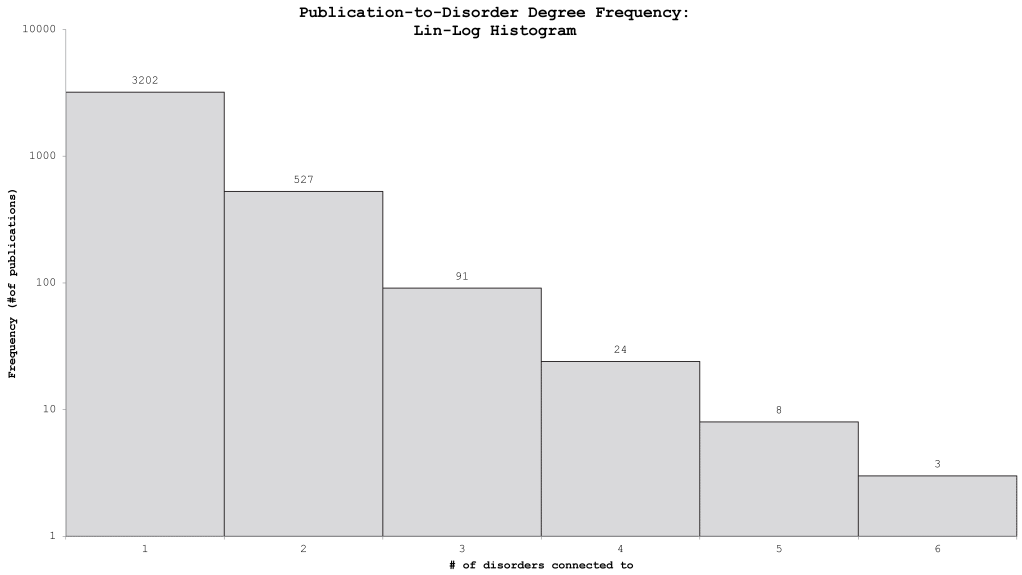

Figure 5. Histogram of disorder-to-publication (logarithmic frequency scale, linear bin scale)

4 Analysis & Discussion

4.1 Match Rate

Of great concern is that our model failed to match 67.75% of abstracts in the corpus from the *Brain Research* journal. At this stage, we can only hypothesize upon the reasons for the lack of matching; a relatively small ontology (with only 260 base terms representing 96 disorders) could be of fault. A widening of the search ontology, in the authors' opinions, is definitely in order. Furthermore, there may exist large number of publications that do not explicitly describe CNS *disorders per se*, but *normal* functioning of the CNS; these publications would therefore evade the Merck classification that was utilized in order to classify the corpus of abstracts. It is of note that further knowledge may be gained by searching for related *science-to-clinic* terminology; one example that was already used in the authors' ontology tree was the association of the terms *nociceptor* and *nociception* with pain disorders. Further discussion of improvements to the ontology is discussed in Section 5 of this report.

4.2 Distorted Distributions & Social Phenomena

Clearly, as delineated prior, the degree distributions of the network we synthesized were not normal (in the parametric statistical sense), as we required non-linear regression in topic ranking studies and/or exponential/reverse-exponential binning in our publication topic distribution models.

Patterns of Sociology Found. Nonetheless, the ideas of the power law, extrapolated by Barabasi[5] and Milojevic[6], must be entertained. The possibility of data bias, i.e., bias inherent in the data set itself and not that imposed by the authors, cannot be excluded.

Classic co-authorship infometrics studies (ones that link papers to authors in similar networks) have shown that there exists` preferential attachment, that is, authors will preferentially attach to other authors who have had prior success publishing.[17] Such co-authorship networks usually show power law (or power law-type) distributions and rankings of node degree, including the Pareto Type II distribution seen in our disorder-to-publication network. It follows, knowing that the disorder is technically a topic, that we can properly speculate there is a preferential attachment of publications (and their authors) to certain topics. Topics of unusually high degree included pain disorders (D = 587 publications), Alzheimer's disorder and tauopathies (D = 408 publications), stroke (D = 406 publications), and anxiety disorders (D = 279 publications). The ranking model (Figure 3) showed solid evidence for this hypothesis. However, the explicit frequency model of the same analysis (Figure 4) showed elements of both a log-normal distribution (with peaks at e^4-e^5 and e^5-e^6 publications) and a long-tailed distribution (with another peak at e^0-e^1 publications per disorder). The significance of the peak for exponents of degrees at 4 shows that there is a concentration of disorder study in more modest disorders, particularly given that this distribution is scaled to linear frequency and not logarithmic.

Nonetheless, viewing the subject from the point of view of those who published the articles we studied, we see that there was a great degree of specialization. As per Figure 5, most publications were connected to only one disorder, coinciding with the first-degree match rates found in Section 3.1 of this paper. From this point, a logarithmic-linear decay occurred, indicating that there may exist disconnects in interdisciplinary studies of separate neuropathologies within the scope of single publications.

5 Future Directions

It is most important to note that while we have reached some conclusions by using the current network, future studies with both this network and its underlying ontology are mandatory. Furthermore, the utilization of this search system as well as graph ontology does not only provide a mapping of disorders to publications; more importantly perhaps, our research forms a framework upon which further clinical and biomedical research may be analyzed.

5.1 Planned Studies: This Network (incl. Improvements to Controlled Vocabulary)

It is very important to note that the ontology (specifically, that of disorder-key term(s)) has not been curated by medical practitioners who deal with the CNS. We wish to subject the aforementioned ontology to validation by a panel of expert clinicians and researchers. Such validation is likely to result in slight modifications to this network based on explicit search terminology used.

After revising the network with the validated ontology, we wish to evaluate the ability of this network to draw conclusions on the interactions of humans with clinical information. We wish to conduct a human survey project that will record the opinions of clinicians and researchers as they pertain to their beliefs on the importance of their own sub-fields of neuroscience and neurology; in this case, our ontological trees represent the sub-fields. Finally, we intend on allowing consumers of healthcare (i.e., the lay public) to interact with this network map and discover how it changes (or reinforces) their perceptions of particular CNS disorders.

Study of Non-pathological CNS Function. It follows that the dichotomy between neurological wellness and illness is becoming blurred, at least if our studies indicate any trends. A high non-match rate between our controlled vocabulary (and its inherent ontology tree) and the abstract corpus, again, supports this idea.

5.2 Ontology as a Framework

With ontology commonly used as a framework in various information science applications, it is clear that this ontology (or a revised version thereof) ought to be used as a framework for the future study of CNS disorders. While we have only applied our ontology to relatively recent articles from *Brain Research,* studies of the resulting network over time (e.g., by comparison to similar networks generated for other publication time periods) would be of great interest. Furthermore, the ontology may be applied outside of *Brain Research* for the purpose of engineering knowledge from any corpus of documents for the purpose of executing any semantic research use case where discovery of knowledge structures is desired.

Creating Frameworks for Consumer Studies and Consumer Applications. As implied in Section 5.1, this ontology may help create a framework specifically for health consumer studies, allowing healthcare consumers to gain knowledge about how the human body is researched in the basic sciences field. It is also possible that such ontology may be useful for semantic analysis of consumers' self-reported health information in order to extract information regarding potential disorders that may be of concern to the consumers and their clinicians.

References

1. Skusa, A., Ruegg, A., Koehler, J.: Extraction of Biological Interaction Networks From Scientific Literature. Briefings in Bioinformatics. 6(3), 264-276 (2005)
2. Spasic I., Ananiadous S., McNaught J., Kumar A. Text Mining and Ontologies in Biomedicine: Making Sense of Raw Text. Briefings in Bioinformatics. 6(3), 239-251 (2005)
3. Yan, E., Ding, Y., Milojevic, S., Sugimoto, C.R.: Topics in Dynamic Research Communities: An Exploratory Story for the Field of Information Retrieval. Journal of Informetrics. 6(1), 140-153 (2012)
4. Newman, M.E.J.: Power Laws, Pareto Distributions, and Zipf's Law. Contemporary Physics 46(5), 323-351 (2005)
5. Barabasi, A.L., Erzsebet, R., Vicsek, T.: Deterministic Scale-Free Networks. Physica Acta. 299, 559-564 (2001)
6. Milojevic, S.: Power-Law Distributions in Information Science – Making the Case for Logarithmic Binning. Journal of the American Society for Information Science and Technology. 61(12), 2417-2425 (2010)
7. Brain Research. (Journal). Information retrieved from http://journals.elsevier.com/brain-research/ on 19 December, 2013.
8. Merck & Co., Inc.: Neurological Disorders. (Section). In The Merck Manual of Diagnosis and Therapy for Professionals. (2013-2014). Retrieved from http://www.merckmanuals.com/professional/neurologic_disorders.html on 09 August 2013.
9. Merck & Co., Inc.: Psychiatric Disorders. (Section). In The Merck Manual of Diagnosis and Therapy for Professionals. (2013-2014) Retrieved from http://www.merckmanuals.com/professional/psychiatric_disorders.html on 09 August 2013.
10. United States Government, National Institutes of Health, National Library of Medicine (n.d.-2014): PubMed (Database Website). Retrieved from http://www.pubmed.gov/ (n.d.)
11. Eclipse Foundation, Inc.: Eclipse IDE for Java EE Developers (Software). Retrieved from http://www.eclipse.org/downloads/
12. Sci2 Team: Sci2 Tool (Software). (2009) Retrieved from http://sci2.cns.iu.edu/
13. Martin, S., Brown, W.M., Klavans, R., Boyack, K.W.: DrL: Distributed Recursive (Graph) Layout. SAND Reports. 2936, 1-10 (2008)
14. Excel 2014 (Software). Microsoft Corporation (Redmond/Seattle, Washington). Retrieved from http://www.microsoft.com/
15. SPSS 20 (Software). International Business Machines (IBM) (Durham, North Carolina). Retrieved from http://www-01.ibm.com/software/analytics/spss/products/statistics/
16. Apache Foundation. Open Office (Software). Retrieved from http://www.openoffice.org/
17. Milojevic, S., Sugimoto, C.R., Yan, E., Ding, Y.: The Cognitive Structure of Library and Information Science: Analysis of Article Title Words. Journal of the American Society for Information Science and Technology. 62(10), 1933-1953 (2011)