Alzheimer's Disease Citation Network Analysis

Peter DeWeirdt

September 18, 2017

Abstract

In this paper we use citation networks to analyze the structure of Alzheimer's Disease research. We use a clustering approach to identify research communities. Then, for each community, we identify research topics using a word correlation network. In our first network we look at Alzheimer's Disease communities in the context of other research communities and note that they share a high number of favorable connections with groups researching cancer. These favorable connections provide grounds for future collaborations between Alzheimer's and cancer researchers. In our second network, we only consider papers focusing on Alzheimer's Disease, and note that this body of research is clearly divided between patient oriented and molecular research. This may be problematic, as we observe that research communities that consist of papers from a wide range of topics tend to have a larger impact upon the collection of Alzheimer's papers.

Background

Current State of Alzheimer's Research

Alzheimer's Disease (AD) is a neurodegenerative disease that causes dementia in patients. AD afflicts an estimated 46.8 million people worldwide along with their families and caregivers [1, 2]. Despite significant advancements in the treatment and diagnosis of AD, no disease-modifying or preventative therapies are known [3]. Innovative research is necessary in order to accelerate the discovery of successful treatments.

In recent years, universities and other research institutions have pushed for increased interdisciplinary research in order to solve complex problems [4]. Recent research by Yegros et al. suggests that interdisciplinary research can lead to knowledge creation or breakthroughs, but can be much riskier and have a higher failure rate than staying safely within a field of research [5]. Thus, it is important to consider which areas of research might be most advantageous for interdisciplinary research.

Citation Networks

We use citation networks to determine significant connections between AD research communities and communities in the wider scientific community. In 1965 Derek de Solla Price recognized that papers and the citations between them could be represented as directed graphs [6]. Since then, citation networks have been used as an important tool in determining the structure of research and

the impact that authors and papers have on a body of research [7]. These types of analyses have become increasingly popular since recent technologies have allowed for easy sharing, storage, and access to millions of scientific papers.

Many real world networks, such as social networks and biological networks have a natural community structure [8]. Citation networks are no exception and can be partitioned into research communities. The directed and time-dependent nature of citation networks makes community detection less straightforward than many other networks. Rosvall and Bergstrom found that random walks on citation networks can mimic information flow, and by minimizing the information flowing between communities, can be used for research community detection [9]. Cheng et al. used the algorithm developed by Rosvall and Bergstrom to resolve the community structure in the Physical Review citation network [10]. They identify significant and robust community structure, and uncover some surprising links across research topics and long stretches of time.

While community structure can be significant in citation networks, it is also important to look at traditional citation metrics for individual papers as well. Some of these metrics include number of citations, impact factor, and h-index. Unfortunately, these metrics only use local information in order to rank papers [7]. To take advantage of the whole network structure, researchers have ranked papers by modeling the diffusion of scientific ideas using Google's PageRank [11] and its variants [12, 13]. We will use both community detection and ranking algorithms to determine the organization of AD research.

Methods

Data Collection

To get the papers for our network, we queried the PubMed database for papers with "Alzheimer's" in the title, abstract, or keywords. Then we extracted the text, year of publication and papers cited for each of these documents. In our first network, we looked at this group of papers (26,464 papers in all) and any paper in the PubMed database within two citations of the original group of papers. This gave us a network of 1,691,139 papers. To make data processing easier we removed papers with less than 8 citations from this network. Then we extracted the largest connected component to obtain a network with 157,958 papers. In our second network we focused on the papers from the original query, thus only targeting those that directly mention AD.

Citation Network

Using our database of papers and the citations between them we created a citation network with the open-source statistical software R [14]. In the citation network, we treated papers as nodes and created a directed edge from paper A to paper B if paper A cites paper B. In order to determine the importance of each paper within the network we calculated each paper's CiteRank centrality [12]. The CiteRank algorithm is a modified version of Google's PageRank [11]. In CiteRank the probability of choosing a paper at random is exponentially proportional to the age of the paper. The CiteRank algorithm mimics the process of a researcher in that a random walker will choose a paper and then with a probability of 0.5, will follow a citation link to another paper also with probability 0.5 the walker will jump to a random paper in the network, once again favoring newer papers to older ones. Each time the walker follows a citation link it transfers weight to the target paper proportional to the weight of the source paper divided by the number of links the source has. In this way, a highly cited paper adds more weight to papers relative to a less cited paper.

We then clustered the papers in the citation network using Infomap [9], one of many available such clustering algorithms. Infomap uses random walks to mimic information flow, and thus works well for

citation networks in which citations tend to indicate the dissemination of ideas in research. In order to determine the significance of our clustering, we compare the modularity of our partition with a randomly rewired network's modularity. Modularity compares the density of edges within a community with the same community in a randomly generated network [15]. In order to measure the density of edges within a community, we take

$$\frac{1}{2}\sum_{i,j} A_{i,j}\delta(c_i, c_j)$$

where $A_{i,j}$ is the adjacency matrix for our network, c_i is the community that node i belongs to, and

$$\delta(c_i, c_j) = \begin{cases} 1 & \text{if } c_i = c_j \\ 0 & \text{otherwise.} \end{cases}$$

Note that we divide by 2 as to avoid double counting edges. Then, in order to find our modularity, Q, we subtract by our random graph model. Thus, we have

$$Q = \sum_{i,j} \frac{1}{2}\left(A_{i,j} - \frac{k_i k_j}{2m}\delta(c_i, c_j)\right).$$

Where k_i is the degree distribution of node i, and m is the number of edges in the network. Note that there is a directed version of modularity [16], but when we calculated it on the network of AD papers, it was nearly identical to the modularity described above. Thus, for simplicity, we only use the non-directed version of modularity.

To determine the robustness of a partition we looked at the normalized mutual information (NMI) between our partition and a the partition from a label propagation algorithm called SpeakEasy [17]. NMI is the measure of similarity between two partitions and is always between 0 and 1, where 1 indicates two identical partitions [18]. Since Infomap and SpeakEasy use vastly different methods to partition a network, a high value for NMI indicates a more robust community structure.

Topic Mapping

To assign papers to research topics, we used the Topic Mapping algorithm by Lancichinetti et. al. [19]. An overview of how the Topic Mapping algorithm works follows, but for more details consult the supplementary materials of the orignial paper. Steps to Topic Mapping:

1. Stem the words of each document, so root words in different forms will be considered the same. For example, "proteins" will be shortened to "protein."
2. Create a matrix of word co-occurences from the stemmed documents. Words are said to co-occur n times if they appear in the title or abstract of n documents together.
3. Compare the co-occurence matrix with a null model. Because each co-occurrence is relatively rare, we can model their frequencies by a Poisson distribution. We are interested in co-occurences that occur at a frequency not expected by random, so we use a p-value of 0.05 in our null model. We then subtract the null model from the co-occurrence matrix and replace any negative value in the resulting matrix with a zero.
4. We use this matrix as a network and cluster words using Infomap.
5. Each cluster is considered a topic, and each paper is assigned to topics by identifying significant overlaps between words in that paper and words assigned to topics. A binomial distribution is used to determine significance of overlap.

6. In order to account for the fact that words do not only belong to one topic, we use PLSA-like likelihood and optimize with LDA likelihood to "blur" topics. This allows very common words, like "disease" to have positive probability of belonging to more than one topic.

We determined the most unique words for each topic, by considering $p(\text{word}|\text{topic})$ and the frequency of the word across all papers. For each topic, we ranked words by their probability and then multiplied this ranking by their frequency. Then, for each topic, we took the ten words with the lowest value from this product. In this way, we chose words that were significant and unique for a given topic. We assigned research communities to the most frequent topic in the community.

Extended Network Results and Discussion

In our extended network, we analyzed the citations between 157,958 papers. The top 5 most important papers, based on CiteRank, can be seen in table 1. We note that these paper's CiteRanks are not directly proportional to the number of citations they have, although well correlated. The first four papers are all DNA or protein sequencing software. The final paper proposes the I^2 metric for measuring the consistencies between the findings of different trails in meta-analyses [20]. Interestingly, although this network started as a core of AD papers, none of the top five most important papers focus on AD.

Table 1: Top 5 most important papers in the Extended network based on CiteRank (CR). Topics can be seen in table 6.

Title	Year	Topic	CR ($\times 10^{-3}$)	# Cites	Ref.
Gapped BLAST and PSI-BLAST: a new generation of protein database search programs	1997	0	3.17	63279	[21]
MEGA5: Molecular Evolutionary Genetics Analysis Using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods	2011	0	3.16	32883	[22]
MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0	2013	0	2.53	17857	[23]
CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice	1994	2	1.93	57707	[24]
Measuring inconsistency in meta-analyses	2003	0	1.87	20123	[20]

We clustered the network using Infomap and obtained 4,524 research communities with $Q = 0.474$. All research communities with a sum of CiteRanks greater than 0.001 can be seen in figure 1. In order to determine the robustness of this result we computed the NMI between this partition and one obtained using SpeakEasy. With the latter partitioning method, we identified 41,836 communities with $Q = 0.182$. Despite these large differences in the number of communities identified and modularity, the two partitions had an NMI of 0.591 indicating a relatively robust partitioning.

Then, in order to determine the significance of the partition, we rewired the edges in the network ($n = 50,000$) while maintaining its degree distribution. While we only rewired less than 3% of the network's edges, the modularity dropped to 0.426, a difference of 0.048. This result tells us that the community structure is significant.

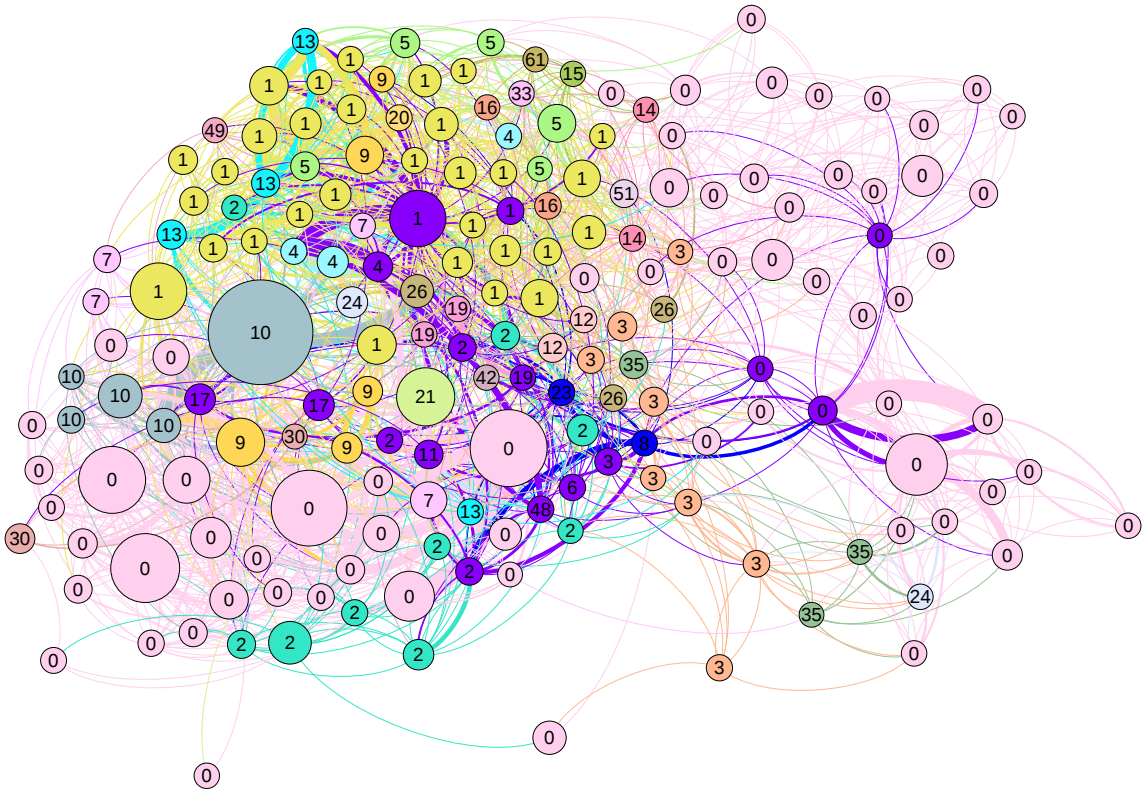


Figure 1: Research communities within the extended citation network. Nodes represent research communities and are sized by the sum of their paper's CiteRanks. Only nodes with a sum of CiteRanks greater than 0.001 are displayed. Each node is labeled by majority topic. Most nodes are colored by majority topic, except nodes assigned to topics 8 or 23 are blue and the communities which they cite are purple. Edges represent citations between communities and are colored by the target community. Edge size is proportional to the number of citations between communities. We only display edges that have a p-value less than 0.001 using a hypergeometric distribution.

We identified 75 topics in the network using Topic Mapping (table 6 in the appendix). An abbreviated version of the topics can be found in table 2. By looking at the most unique words for each topic, we determined that topics 8 and 23 were most clearly related with AD. Papers assigned to topic 8 focus on the aβ protein, while papers assigned to topic 23 focus on the tau protein. Note that no distinct topic was identified for AD patient studies specifically, as these papers likely were assigned to topic 0, which encompasses a wide range of patient association studies. We found 68 and 15 research communities which had topics 8 and 23 as their most frequent research topic

respectively. In order to determine the importance of the relationship between two topics, say T_1 and T_2, we defined the following metric

$$\text{Contribution of } T_2 \text{ to } T_1 = \sum_{c_i \in T_1} \sum_{c_j \in T_2} \frac{\text{\# of significant citations from } c_i \text{ to } c_j}{\text{\# of significant citations from } c_i} \cdot CiteRank_{c_i}. \quad (1)$$

Where c_i and c_j are research communities. A group of citations from one community to another is considered significant if it has a p-value less than 0.001 when considering a hypergeometric distribution. We normalize this value by the sum of every topic's contribution to T_1. This means that values for contribution will always add up to 1 for a specific topic. The top 10 most valued contributions for topics 8 and 23 can be seen in table 3.

Table 2: Top 10 most unique words for select topics in the extended network identified with Topic Mapping. Papers refers to the percent of papers belonging to a given topic. A full version of the extended network topic table can be found in the appendix.

Topic Num.	Top Words	Papers (%)
0	studi, associ, cognit, patient, use, risk, method, age, data, diseas	23.07
1	tumor, cell, cancer, signal, nfκb, activ, t, express, prolifer, apoptosi	9.33
2	structur, membran, domain, vesicl, conform, fold, protein, residu, peptid, golgi	8.98
3	receptor, channel, ca2, synapt, glutam, calcium, neuron, nmda, synaps, nachr	6.31
4	mitochondril, ros, mitochondri, oxid, oxygen, antioxid, redox, reactiv, superoxid, nadph	3.43
8	aβ, app, aβ42, plaqu, ad, deposit, bace1, γsecretas, alzheim, presenilin	2.17
9	methyl, histon, deacetylas, epigenet, acetyl, chromatin, hdac, cpg, h3, sirt1	2.13
10	mirna, rnas, translat, noncod, rna, mrnas, ribosom, microrna, gene, posttranscript	2.80
11	prion, prp, prpc, prpsc, scrapi, psi, spongiform, cjd, creutzfeldtjakob, bse	0.82
12	retin, retina, amd, photoreceptor, rpe, eye, degener, macular, pigment, glaucoma	1.05
17	splice, exon, intron, premrna, smn, sma, hnrnp, altern, sr, a1	1.12
19	hd, huntington, htt, polyglutamin, huntingtin, repeat, polyq, cag, ataxia, c9orf72	1.18
23	tau, neurofibrillari, tangl, hyperphosphoryl, tauopathi, cdk5, microtubuleassoci, neurofila, gsk3β, tdp43	1.04
48	pd, lrrk2, αsynuclein, dopaminerg, parkinson, αsyn, nigra, substantia, pink1, parkin	0.81

Table 3: Top 10 most valuable topics for research communities with majority topics 8 (3a) and 23 (3b). The value for each topic is measured by (1) and is normalized to sum to one. Topics can be seen in table 2.

(a) Most Valued Topics for Topic 8 Research Communities.

Rank	Topic	Value
1	8	0.28
2	2	0.19
3	0	0.14
4	1	0.13
5	23	0.05
6	10	0.03
7	11	0.02
8	3	0.01
9	12	0.01
10	9	0.01

(b) Most Valued Topics for Topic 23 Research Communities.

Rank	Topic	Value
1	1	0.35
2	0	0.25
3	2	0.09
4	23	0.08
5	48	0.03
6	8	0.03
7	4	0.03
8	19	0.02
9	11	0.02
10	17	0.02

We were surprised to find that topic 23 research communities get more value citing outside topics than citing other topic 23 communities. Thus, we can consider topic 23 a secondary research area, supporting other areas of research, but not necessarily having strong intra-topic ties. We found the strong connections that both topics had with topic 1 to be the most surprising. It contributes a value of 0.13 to topic 8 and a value of 0.35 to topic 23. We can see from table 2 that papers assigned to topic 1 focus on the cellular origins of cancer. Past studies have noted an inverse relationship between cancer and AD in patients [25], as well as the possibility of a common biological mechanism for the origin of both diseases [26]. Given the value that cancer research has added to AD research communities, as well as the potential biological connections between the two diseases, we recommend increased collaboration between AD and cancer researchers.

Alzheimer's Network Results and Discussion

We determined the CiteRank for each paper in the Alzheimer's network by running the CiteRank algorithm on the 24,402 Alzheimer's papers. The papers with the highest value for CiteRank can be seen in table 4. We can see that the papers span a 26 year range and cover three distinct topics. These five papers account for over 1.4% of the total CiteRank while simultaneously representing less than 0.01% of papers in our network. Interestingly, the National Institute on Aging-Alzheimer's Association has two papers in the top 5, thus having a disproportionately large impact on the body of AD research.

In the network we found 1244 research communities using Infomap, giving us a modularity of 0.359. This compares with an modularity of 0.187 for a rewired network (n = 25,000) that maintained the same degree distribution. Thus, we can say the community structure in the AD network is significant. We also observed an NMI of 0.756 with SpeakEasy's partitioning of the network, indicating a robust clustering.

We identified 22 topics using the Topic Mapping algorithm, which can be seen in table 7 in the appendix. We found that more than **68%** of all contributions in the network, as measured by

Table 4: Top 5 most important papers in the AD network based on CiteRank (CR). Topics can be seen in table 7 in the appendix.

Title	Year	Topic	CR ($\times 10^{-3}$)	# Cites	Ref.
Amyloid plaque core protein in Alzheimer disease and Down syndrome	1985	1	2.90	4206	[27]
The diagnosis of dementia due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease	2011	0	2.30	4515	[28]
Diffusible, nonfibrillar ligands derived from Aβ1–42 are potent central nervous system neurotoxins	1998	1	2.00	3283	[29]
Apolipoprotein E: high-avidity binding to beta-amyloid and increased frequency of type 4 allele in late-onset familial Alzheimer disease	1993	14	1.88	4125	[30]
Toward defining the preclinical stages of Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease	2011	0	1.70	2950	[31]

CiteRank, came from topics 0 and 1 (table 5). Papers assigned to topic 0 appear to use patient data in their analyses, while papers assigned to topic 1 focus on the molecular nature of AD, concentrating on aβ and tau proteins. Visually inspecting the citation network using an attraction-repulsion algorithm [32] we can see a clear division between research communities focused on these topics (figure 2).

We suspect that the divide between these two areas of AD research are slowing drug development and diagnosis improvement. Researchers focusing on patients must also consider the molecular nature of the disease, just as researchers focusing on the molecular causes of AD must also keep the patient in mind. A connection should be made between both of these sides of AD research in order to uncover novel therapies.

From our network, we isolated the four communities which had the two highest and two lowest values for the geometric mean of CiteRanks (figure 3). These communities indicate that a diversity of topics within a research community can increase its mean CiteRank. In fact, we observed a significant correlation (p-value 3e-04) of 0.1 between the number of topics in a community and its mean CiteRank. Interestingly, we did not find any significant correlation between how well connected a community was, as measured by its number of intra-community citations divided by the number of possible intra-community citations, and its mean CiteRank. This likely indicates that research communities that are not well connected within the community, are well connected

Table 5: Top 10 most unique words for select topics in the AD network identified with Topic Mapping. Papers refers to the percent of papers belonging to a given topic. A full version of the topic table can be found in the appendix.

Topic Num.	Top Words	Papers (%)
0	ad, cognit, dementia, patient, subject, clinic, mci, group, mild, diseas	33.46
1	protein, aβ, app, cell, peptid, aggreg, activ, neuron, tau, inhibit	34.71

with other nodes in the network.

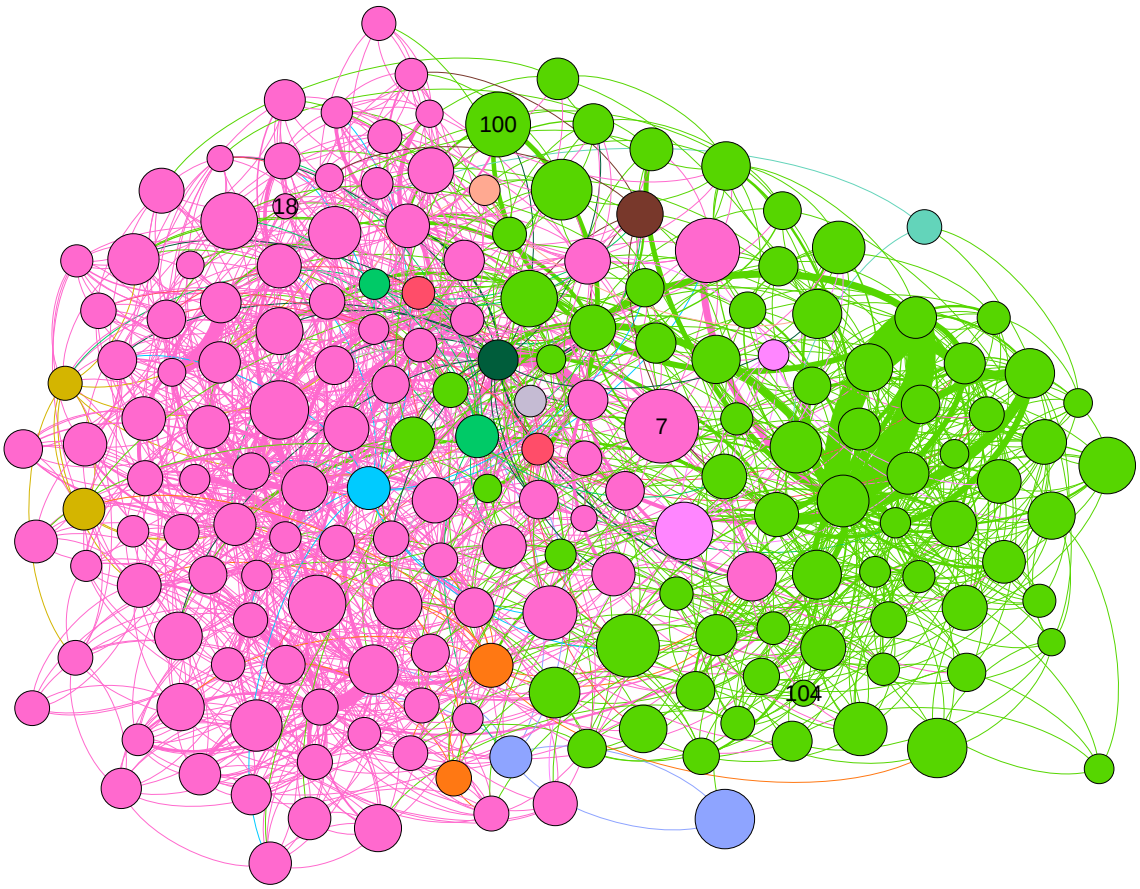
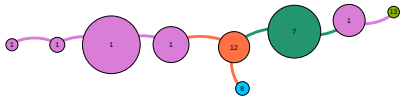
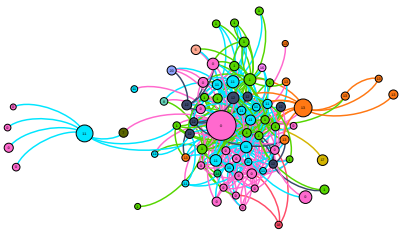


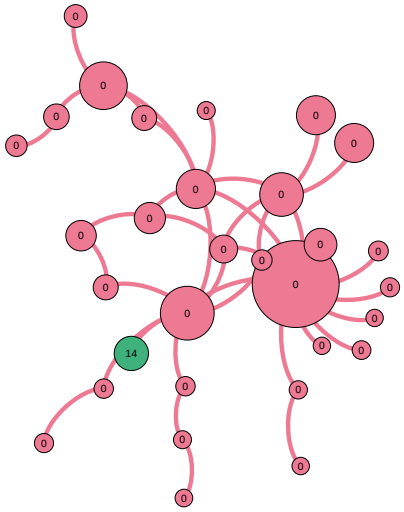
Figure 2: Research communities within the AD citation network. Nodes represent research communities and are sized by the geometric mean of its paper's CiteRank. Only nodes with a sum of cite ranks greater than 0.001 are displayed. Each node is colored by majority topic, which can be seen in full in table 7 in the appendix. Green nodes belong to topic 0, while pink nodes belong to topic 1. Labeled nodes are displayed in figure 3. Edges represent citations between communities and are colored by the target community. Edge size is proportional to the number of citations between communities. We only display edges that have a p-value less than 0.001 using a hypergeometric distribution.



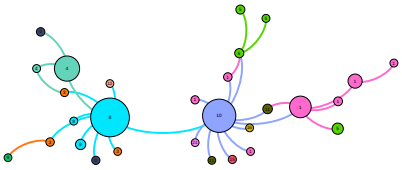
(a) Research community 7. The community has a mean CiteRank of 6.6E-5. Topic 1 is the majority topic for this community.



(b) Research community 100. The community has a mean CiteRank of 5.6E-5. Topic 0 is the majority topic for this community.



(c) Research community 104. The community has a mean CiteRank of 1.5E-5. Topic 0 is the majority topic for this community.



(d) Research community 18. The community has a mean CiteRank of 1.4E-5. Topic 1 is the majority topic for this community.

Figure 3: Research communities from the AD citation network in figure 2. Figures 3a and 3b, and 3c and 3d represent the two highest and two lowest values for mean CiteRank respectively. Nodes represent individual papers and are labeled by topic, all of which can be seen in table 7. Note that the color of each node does not correspond with the colors in table 7. Nodes are sized by CiteRank. Edges represent citations and are colored by which paper is being cited.

Considerations and Future Research

For data collection, we note that PubMed is limited in the number of papers within its database. Other databases, like Web of Science might yield more accurate results. Furthermore, discounting papers with less than 8 citations in the extended network may have skewed results. Westergaard et al. note that just using the title and abstract to label papers might yield different results than text mining the whole paper [33]. In future research we hope to take these considerations into account in order to obtain a more representative network.

While communities appeared to be significant and robust, more rigorous methods should be implemented to analyze them. Assigning each community to its most frequent topic is problematic for research communities that cover a wide range of topics. Furthermore, trying to determine the success of each research community using the geometric mean of its paper's CiteRanks is likely not the most precise measure of a community's success. Using this measure, smaller communities have a much higher chance of randomly being successful when compared with larger communities. Not only was intra-comunnity value hard to determine, but also inter-community value was difficult to quantify. The equation we obtained (1) only represents one attempt at quantifying the relationship between communities. A rigorous analysis of this approach is required in order for the methodology to be completely reliable.

Nonetheless, the results we found show promise for future research. We noted that diversity of research had a positive correlation with the geometric mean of AD communities. We would like to look at what other factors make research communities successful. We also noted a clear divide in AD research. It would seem important to identify the communities that are most successfully bridging the two research areas. Most importantly, we would like to determine which communities are helping to discover novel and promising therapies for AD. Taking the above considerations into account, citation networks show promise for helping to accelerate the research process.

References

[1] J. Cummings, P. S. Aisen, B. DuBois, L. Frölich, C. R. Jack, R. W. Jones, J. C. Morris, J. Raskin, S. A. Dowsett, and P. Scheltens, "Drug development in alzheimer's disease: the path to 2025," *Alzheimer's research & therapy*, vol. 8, no. 1, p. 39, 2016.

[2] A. Association *et al.*, "2016 alzheimer's disease facts and figures," *Alzheimer's & Dementia*, vol. 12, no. 4, pp. 459–509, 2016.

[3] H. Hampel, D. Prvulovic, S. Teipel, F. Jessen, C. Luckhaus, L. Frölich, M. W. Riepe, R. Dodel, T. Leyhe, L. Bertram, *et al.*, "The future of alzheimer's disease: the next 10 years," *Progress in neurobiology*, vol. 95, no. 4, pp. 718–728, 2011.

[4] J. A. Jacobs and S. Frickel, "Interdisciplinarity: A critical assessment," *Annual review of Sociology*, vol. 35, pp. 43–65, 2009.

[5] A. Yegros-Yegros, I. Rafols, and P. D'Este, "Does interdisciplinary research lead to higher citation impact? the different effect of proximal and distal interdisciplinarity," *PloS one*, vol. 10, no. 8, p. e0135095, 2015.

[6] D. J. D. S. Price, "Networks of scientific papers," *Science*, pp. 510–515, 1965.

[7] F. Radicchi, S. Fortunato, and A. Vespignani, "Citation networks," in *Models of science dynamics*, pp. 233–257, Springer, 2012.

[8] M. A. Porter, J.-P. Onnela, and P. J. Mucha, "Communities in networks," *Notices of the AMS*, vol. 56, no. 9, pp. 1082–1097, 2009.

[9] M. Rosvall and C. T. Bergstrom, "Maps of random walks on complex networks reveal community structure," *Proceedings of the National Academy of Sciences*, vol. 105, no. 4, pp. 1118–1123, 2008.

[10] P. Chen and S. Redner, "Community structure of the physical review citation network," *Journal of Informetrics*, vol. 4, no. 3, pp. 278–290, 2010.

[11] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web.," tech. rep., Stanford InfoLab, 1999.

[12] D. Walker, H. Xie, K.-K. Yan, and S. Maslov, "Ranking scientific publications using a model of network traffic," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2007, no. 06, p. P06010, 2007.

[13] P. Chen, H. Xie, S. Maslov, and S. Redner, "Finding scientific gems with google's pagerank algorithm," *Journal of Informetrics*, vol. 1, no. 1, pp. 8–15, 2007.

[14] R. C. Team, "R language definition," *Vienna, Austria: R foundation for statistical computing*, 2000.

[15] M. E. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Physical review E*, vol. 69, no. 2, p. 026113, 2004.

[16] E. A. Leicht and M. E. Newman, "Community structure in directed networks," *Physical review letters*, vol. 100, no. 11, p. 118703, 2008.

[17] C. Gaiteri, M. Chen, B. Szymanski, K. Kuzmin, J. Xie, C. Lee, T. Blanche, E. C. Neto, S.-C. Huang, T. Grabowski, *et al.*, "Identifying robust communities and multi-community nodes by combining top-down and bottom-up approaches to clustering," *Scientific reports*, vol. 5, 2015.

[18] A. Lancichinetti, S. Fortunato, and J. Kertész, "Detecting the overlapping and hierarchical community structure in complex networks," *New Journal of Physics*, vol. 11, no. 3, p. 033015, 2009.

[19] A. Lancichinetti, M. I. Sirer, J. X. Wang, D. Acuna, K. Körding, and L. A. N. Amaral, "High-reproducibility and high-accuracy method for automated topic classification," *Physical Review X*, vol. 5, no. 1, p. 011007, 2015.

[20] J. P. Higgins, S. G. Thompson, J. J. Deeks, and D. G. Altman, "Measuring inconsistency in meta-analyses," *BMJ: British Medical Journal*, vol. 327, no. 7414, p. 557, 2003.

[21] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped blast and psi-blast: a new generation of protein database search programs," *Nucleic acids research*, vol. 25, no. 17, pp. 3389–3402, 1997.

[22] K. Tamura, D. Peterson, N. Peterson, G. Stecher, M. Nei, and S. Kumar, "Mega5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods," *Molecular biology and evolution*, vol. 28, no. 10, pp. 2731–2739, 2011.

[23] K. Tamura, G. Stecher, D. Peterson, A. Filipski, and S. Kumar, "Mega6: molecular evolutionary genetics analysis version 6.0," *Molecular biology and evolution*, vol. 30, no. 12, pp. 2725–2729, 2013.

[24] J. D. Thompson, D. G. Higgins, and T. J. Gibson, "Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice," *Nucleic acids research*, vol. 22, no. 22, pp. 4673–4680, 1994.

[25] J. A. Driver, A. Beiser, R. Au, B. E. Kreger, G. L. Splansky, T. Kurth, D. P. Kiel, K. P. Lu, S. Seshadri, and P. A. Wolf, "Inverse association between cancer and alzheimer's disease: results from the framingham heart study," *Bmj*, vol. 344, p. e1442, 2012.

[26] M. I. Behrens, C. Lendon, and C. M. Roe, "A common biological mechanism in cancer and alzheimer's disease?," *Current Alzheimer Research*, vol. 6, no. 3, pp. 196–204, 2009.

[27] C. L. Masters, G. Simms, N. A. Weinman, G. Multhaup, B. L. McDonald, and K. Beyreuther, "Amyloid plaque core protein in alzheimer disease and down syndrome," *Proceedings of the National Academy of Sciences*, vol. 82, no. 12, pp. 4245–4249, 1985.

[28] G. M. McKhann, D. S. Knopman, H. Chertkow, B. T. Hyman, C. R. Jack, C. H. Kawas, W. E. Klunk, W. J. Koroshetz, J. J. Manly, R. Mayeux, *et al.*, "The diagnosis of dementia due to alzheimer's disease: Recommendations from the national institute on aging-alzheimer's association workgroups on diagnostic guidelines for alzheimer's disease," *Alzheimer's & dementia*, vol. 7, no. 3, pp. 263–269, 2011.

[29] M. P. Lambert, A. Barlow, B. A. Chromy, C. Edwards, R. Freed, M. Liosatos, T. Morgan, I. Rozovsky, B. Trommer, K. L. Viola, *et al.*, "Diffusible, nonfibrillar ligands derived from aβ1–42 are potent central nervous system neurotoxins," *Proceedings of the National Academy of Sciences*, vol. 95, no. 11, pp. 6448–6453, 1998.

[30] W. J. Strittmatter, A. M. Saunders, D. Schmechel, M. Pericak-Vance, J. Enghild, G. S. Salvesen, and A. D. Roses, "Apolipoprotein e: high-avidity binding to beta-amyloid and increased frequency of type 4 allele in late-onset familial alzheimer disease.," *Proceedings of the National Academy of Sciences*, vol. 90, no. 5, pp. 1977–1981, 1993.

[31] R. A. Sperling, P. S. Aisen, L. A. Beckett, D. A. Bennett, S. Craft, A. M. Fagan, T. Iwatsubo, C. R. Jack, J. Kaye, T. J. Montine, *et al.*, "Toward defining the preclinical stages of alzheimer's disease: Recommendations from the national institute on aging-alzheimer's association workgroups on diagnostic guidelines for alzheimer's disease," *Alzheimer's & dementia*, vol. 7, no. 3, pp. 280–292, 2011.

[32] Y. Hu, "Efficient, high-quality force-directed graph drawing," *Mathematica Journal*, vol. 10, no. 1, pp. 37–71, 2005.

[33] D. Westergaard, H.-H. Stærfeldt, C. Tønsberg, L. J. Jensen, and S. Brunak, "Text mining of 15 million full-text scientific articles," *bioRxiv*, p. 162099, 2017.

Appendix

Topic	Top Unique Words	Papers (%)
0	studi, associ, cognit, patient, use	23.07
1	tumor, cell, cancer, signal, nfκb	9.33
2	structur, membran, domain, vesicl, conform	8.98
3	receptor, channel, ca2, synapt, glutam	6.31
4	mitochondril, ros, mitochondri, oxid, oxygen	3.43
5	insulin, glucos, obes, metabol, adipos	2.27
6	virus, hiv1, viral, hiv, infect	1.83
7	repair, telomer, dna, break, atm	1.46
8	aβ, app, aβ42, plaqu, ad	2.17
9	methyl, histon, deacetylas, epigenet, acetyl	2.13
10	mirna, rnas, translat, noncod, rna	2.80
11	prion, prp, prpc, prpsc, scrapi	0.82
12	retin, retina, amd, photoreceptor, rpe	1.05
13	elegan, drosophila, lifespan, caenorhabd, longev	1.50
14	dha, fatti, n3, aa, pufa	0.76
15	lipoprotein, cholesterol, apo, ldl, hdl	0.86
16	barrier, junction, bbb, bloodbrain, permeabl	1.29
17	splice, exon, intron, premrna, smn	1.12
18	estrogen, steroid, hormon, e2, erα	0.83
19	hd, huntington, htt, polyglutamin, huntingtin	1.18
20	skelet, muscl, myoblast, exercis, myogen	0.78
21	intestin, gut, microbiota, periodont, immun	0.84
22	ferritin, iron, hepcidin, transferrin, heme	0.42
23	tau, neurofibrillari, tangl, hyperphosphoryl, tauopathi	1.04
24	circadian, sleep, clock, rhythm, oscil	0.77
25	hypertens, ii, angiotensin, ang, pressur	0.36
26	deliveri, nanoparticl, drug, exosom, target	2.17
27	x, fragil, fmr1, fxs, fmrp	0.42
28	kidney, renal, glomerular, klotho, podocyt	0.38
29	plant, arabidopsi, thaliana, tobacco, coexpress	0.67
30	coli, escherichia, biofilm, salmonella, typhimurium	0.32
31	malaria, parasit, gondii, toxoplasma, plasmodium	0.55
32	sulfat, proteoglycan, heparan, hs, heparin	0.59
33	complement, c3, c1q, c5a, sle	0.61
34	ceramid, sphingolipid, sphingosin, sphingomyelinas, asm	0.53
35	olfactori, neurogenesi, odor, bulb, svz	1.76
36	oglcnac, glycosyl, oglcnacyl, ogt, olink	0.49
37	toxin, venom, neurotoxin, cholera, ct	0.18
38	h2s, homocystein, cbs, hci, vascular	0.46
39	zinc, metal, mt, zn2, metallothionein	0.20
40	cu, copper, atp7b, atp7a, wilson	0.17

Continued on next page

Topic	Top Unique Words	Papers (%)
41	chlamydia, pneumonia, chlamydi, eb, trachomati	0.20
42	sperm, fertil, germ, spermatogenesi, testi	0.34
43	cytokin, lps, tryptophan, kynurenin, ido	0.51
44	acetylcholinesteras, ginseng, compound, extract, ach	0.19
45	astrocyt, cord, spinal, gfap, aqp4	0.81
46	camp, pka, phosphodiesteras, cyclic, cgmp	0.35
47	pgp, pglycoprotein, mdr, multidrug, paclitaxel	0.22
48	pd, lrrk2, αsynuclein, dopaminerg, parkinson	0.81
49	ubiquitin, ligas, sumoyl, e3, sumo	0.70
50	la, fear, en, el, syn	0.20
51	rage, glycat, hmgb1, diabet, srage	0.30
52	dynein, cilia, kinesin, shh, ciliari	0.44
53	nitrat, phi, usepackageamsmath, usepackagewasysym, usepackageamsfont	0.08
54	car, p450, pb, cyp, cyp3a4	0.17
55	mecp2, bdnf, rett, ds, ipsc	0.83
56	spindl, mitosi, mitot, chromosom, divis	0.41
57	thyroid, tr, tsh, t3, trh	0.11
58	resveratrol, polyphenol, egcg, curcumin, tea	0.43
59	airway, cftr, fibrosi, cystic, cf	0.38
60	stress, amygdala, behavior, cocain, ethanol	1.13
61	peroxisom, pparγ, proliferatoractiv, ppar, autophag	0.34
62	beclin, selenium, se, selenoprotein, sec	0.07
63	ra, vitamin, retino, retinoid, atra	0.13
64	lysosom, atp13a2, palmitoyl, infantil, cathepsin	0.29
65	eno, nos, ino, nitric, endotheli	0.18
66	len, cataract, αbcrystallin, crystallin, fiber	0.14
67	lamin, b1, senesc, nuclear, farnesyl	0.23
68	tyrosin, src, guidanc, fak, signal	1.17
69	proteincoupl, gpcrs, platelet, lpa, gpcr	0.35
70	pheromon, cerevisia, mate, saccharomyc, calcineurin	0.40
71	nrf2, ahr, nqo1, aryl, hydrocarbon	0.18
72	gangliosid, mag, gm1, sialic, glycosphingolipid	0.15
73	myelin, schizophrenia, schwann, oligodendrocyt, disc1	0.40
74	bryostatin, sema5a, glanc, NA, scg10	0.45

Table 6: Top 5 most unique words for each topic in the extended network identified with Topic Mapping. Papers refers to the percent of papers belonging to a given topic.

Topic	Top Unique Words	Papers (%)
0	ad, cognit, dementia, patient, subject	33.46
1	protein, aβ, app, cell, peptid	34.71
2	ach, acetylcholinesteras, inhibitor, medicin, effect	2.64
3	cholinerg, ngf, forebrain, neuron, p75ntr	2.42
4	pd, αsynuclein, lewi, αsyn, parkinson	1.82

Continued on next page

Topic	Top Unique Words	Papers (%)
5	insulin, diet, diabet, mice, t2dm	3.46
6	nicotin, nachr, receptor, α7, agonist	1.30
7	ds, chromosom, syndrom, ts65dn, trisomi	1.42
8	retin, layer, optic, retina, glaucoma	2.07
9	estrogen, hormon, e2, steroid, postmenopaus	1.11
10	acid, dha, fatti, aa, lipid	0.93
11	la, trem2, de, novo, pgrn	0.44
12	materi, supplementari, user, onlin, author	1.24
13	gene, mirna, microrna, rna, transcriptom	2.81
14	apo, apoe4, allel, apolipoprotein, apoe3	1.42
15	dna, methyl, epigenet, histon, repair	1.48
16	sleep, circadian, rhythm, melatonin, oscil	1.13
17	iron, aluminum, copper, metal, zinc	1.50
18	tdp43, repeat, expans, c9orf72, mutat	0.77
19	autism, seizur, fragil, x, pdapp	0.80
20	tbi, injuri, traumat, chronic, encephalopathi	2.31
21	metabolit, caffein, metabolom, kynurenin, coffe	0.77

Table 7: Top 5 most unique words for each topic in the AD network identified with Topic Mapping. Papers refers to the percent of papers belonging to a given topic.