All Roads Lead to Philosophy

Examining the 'Getting to Philosophy' Phenomena on Wikipedia using Network Analysis

Austin Barish

8/7/23

Table of contents

1 Abstract **2**

2 Introduction **2**

3 Methods **6**
3.0.1 Finding First Links . . . 6
3.0.2 Creating the First-Link Network . . . 7
3.0.3 Creating the Second-Link Network . . . 9
3.0.4 Plotting Methods . . . 10

4 Results **10**
4.1 First Link Network . . . 10
4.1.1 Convergence . . . 10
4.1.2 Notable Nodes and Paths to Philosophy . . . 12
4.1.3 First-Link Network Structure . . . 18
4.1.4 Degree vs. Distance from the Philosophy Page . . . 18
4.2 Second Link Network . . . 18
4.2.1 Second Link Convergence . . . 18
4.2.2 Second Link Largest Connected Components . . . 21
4.2.3 Second Link Notable Nodes . . . 21
4.2.4 Second Link Network Structure . . . 25

5 Conclusions **25**
5.1 First-Link Network Conclusions . . . 25
5.1.1 Getting to Philosophy . . . 25
5.1.2 Why is the Philosophy Node Significant and What is the Nature of the Most Common First Links? . . . 26

5.1.3 Distance from Philosophy and Degree 27
5.2 Second-Link Network Conclusions 27
5.3 References 27

Keywords

- Wikipedia
- Philosophy
- Getting to Philosophy
- First Page Philosophy
- First Link
- Navigability

1 Abstract

In this study, I analyze a phenomenon on Wikipedia in which repeatedly clicking "first link" of a webpage invariably takes a user to the Philosophy page. I examine the percent of pages on Wikipedia in which this idea holds true in an effort to understand how Wikipedia's network is structured and what that means for its user navigability and understanding. Previous research indicates that users' page navigation is heavily focused on the lead of a Wikipedia article, rarely venturing beyond the first paragraph(Dimitar Dimitrov 2016); therefore, I limit my analysis to the first several links in this section; further analysis with greater computing power could be done on the links within the entire article. Amongst these first several links, I seek to determine if there are any other link locations that reach a specific page with any abnormal frequencies, including the philosophy page. To conduct my analysis, I construct a network using Wikipedia pages as nodes and the links on the page as directed links between nodes. I collected my data using a Breadth-First Search (BFS), meaning once I reach a page that has already been determined to reach the philosophy page, I move on to another root page. With the network, I examine average path lengths to the philosophy page, the neighbors of the philosophy page that most commonly direct to it, and the structure of the first-link network itself. Furthermore, I examine the second link of Wikipedia pages and conduct and analysis of that network as well. My conclusions demonstrate the effectiveness of Wikipedia's effort to make their introductory sentence and links sufficiently broad.

2 Introduction

Wikipedia pages are built with the user's understanding in mind. To ensure consistency across pages and maintain reliability as a credible source, there are extensive guidelines on the structure of each page. As one of the most important components of a Wikipedia page, linked content and the content of the lead paragraph is tightly monitored. Links serve to "provide instant pathways to locations within and outside the project that can increase readers'

understanding of the topic at hand."(Wikipedia 2023c) Users will click on links when a topic is unfamiliar to them, or if they interested in learning more.

When arriving to a page, a user ought to have the topic explained to them as though they know little to nothing about it. The lead ought to frame the reader so as to "set the scene of the topic".(Wikipedia 2023c) Wikipedia explains the structure of the lead paragraph:

> In Wikipedia, the lead section is an introduction to an article and a summary of its most important contents. It is located at the beginning of the article, before the table of contents and the first heading. It is not a news-style lead or "lede" paragraph.
>
> The average Wikipedia visit is a few minutes long. The lead is the first thing most people will read upon arriving at an article, and may be the only portion of the article that they read. It gives the basics in a nutshell and cultivates interest in reading on—though not by teasing the reader or hinting at what follows. It should be written in a clear, accessible style with a neutral point of view.(Wikipedia 2023c)

Wikipedia goes on to outline how the opening paragraph and sentence ought to be structured. They explain that the "[The opening paragraph] should establish the context in which the topic is being considered by supplying the set of circumstances or facts that surround it. If appropriate, it should give the location and time."(Wikipedia 2023c) For example, a building's first link will most likely be its location. Within that paragraph, its opening sentence is critical for my study as it will contain the first link. Editors are instructed that "the first sentence should tell the nonspecialist reader what or who the subject is, and often when or where."(Wikipedia 2023c) They go on to provide explicit instructions on what the first linked topic ought to be in an article:

> The first sentence should provide links to the broader or more elementary topics that are important to the article's topic or place it into the context where it is notable.
>
> For example, an article about a building or location should include a link to the broader geographical area of which it is a part.
>
> Arugam Bay is a bay on the Indian Ocean in the dry zone of Sri Lanka's southeast coast.
>
> In an article about a technical or jargon term, the first sentence or paragraph should normally contain a link to the field of study that the term comes from.
>
> In heraldry, tinctures are the colours used to emblazon a coat of arms.
>
> The first sentence of an article about a person should link to the page or pages about the topic where the person achieved prominence.

> *Harvey Lavan "Van" Cliburn Jr.* (July 12, 1934 – February 27, 2013) was an American pianist who achieved worldwide recognition in 1958 at age 23, when he won the first quadrennial International Tchaikovsky Piano Competition in Moscow, at the height of the Cold War.
>
> Exactly what provides the context needed to understand a given topic varies greatly from topic to topic.(Wikipedia 2023c)

As you can see, the first link of each page will be increasingly broad as you continue to click the first link. These instructions create a picture of how a topic like philosophy can be at the center of Wikipedia's first link network. Conversely, it is doubtful that such a center exists for any other link placement. Even just the second link in an article can be increasingly specific, moving laterally or even backwards in specificity rather than towards larger hubs such as philosophy. Take one of Wikipedia's examples in Harvey Lavan "Van" Cliburn Jr; his first link path begins with pianist then continues as follows: piano, keyboard instrument, musical instrument, music, art, creativity, psychology, mind, thought, consciousness, awareness, philosophy. With each passing link you can sense that your destiny on the philosophy page grows closer; the topics are broader and the connection from it to philosophy feels increasingly obvious. However, if we were to follow the second link, International Tchaikovsky Piano Competition, we find ourselves on the following path: Saint Petersburg, Russia, Eastern Europe, Ural Mountains, Eurasia, Europe, peninsulas, mainland, continent, regions, Earth's surface, hemispheres, etc. Unlike with the first link, the second link gets stuck in geographic limbo without ever getting closer to a central topic like Philosophy. I will explore what a second link network looks like further in my analysis and see that geography, broadly speaking, is the typical destination of pages when clicking the second-link.

There is special focus on the very beginning of a Wikipedia page because that is where users devote most of their attention. Dimitrov et al. utilize click data from Wikipedia's navigation logs to construct a heat map of where users are clicking the most on Wikipedia pages. The heat map illustrates two clear dark red, high density, lines at the beginning of the page directly where the lead is located, demonstrating that users highest click rate is on links within the first few lines of the opening paragraph. The rest of the page is sparse beyond a preference for links on the left side of pages, a phenomenon the authors themselves do not fully understand.(Dimitar Dimitrov 2016) However, the high click rate within the lead indicates to us that understanding the nature of the network of the first few links in an article is indicative of the nature of the network that users are typically interacting with.

Research has already been done into the size of the Giant Connected Component (GCC) of nodes that connect to the philosophy node. In a study of Wikipedia's navigability by language, as of 2017, 97.0% of pages in English will reach the philosophy page(Daniel Lamprecht 2016), a slight increase of around 2.5% since 2011.(Wikipedia 2023b) These numbers fluctuate across languages, with some languages have a center on pages such as Psychology in Spanish or Person in Japanese each with varying sizes but still having the majority of nodes reach these pages(Daniel Lamprecht 2016); my study will only be focused on the English network of

Wikipedia. In the future, it would be interesting to study this phenomenon in other languages as I have done with English. In particular, previous studies indicate that Dutch has the smallest GCC with just 67.0% of nodes in its GCC.(Daniel Lamprecht 2016) I would like to compare its network to English to understand this discrepancy. However, the English network is already far large enough for the scope of this study.

If you would like to see how this network is formed beyond clicking through Wikipedia webpages on your own, the online page xefer will quickly build out a network of pages and their first links until you reach the philosophy page. This is a helpful tool that is good to visualize what this can look like in practice. However, it was designed to always reach the philosophy page even for those pages that manage to avoid the philosophy page. It does this by skipping to the second link on a page when it realizes it will not be able to reach the philosophy page through the first link.(xefer 2011) Therefore, we need to construct our own network to understand these disconnected nodes.

To understand how a node can be disconnected, we ought to look at what makes philosophy the center of the network. If you click on the first link on the philosophy page, you will find yourself back on the philosophy page in 6 clicks. This self-loop forms a bottom of sorts to the network as nothing beyond the 5 pages you reach from the philosophy page can be found from there. Amongst those 5 pages, philosophy is by far the largest and most central node, making it the logical choice for the center. For another node to avoid the philosophy node, it would require a similar cycle. Therefore, it is going to be a broad topic as it has to be something that could similarly be in the first sentence of a Wikipedia page. This eliminates super specific pages from consideration despite them being the intuitive guess for what might manage to avoid philosophy. However, these specific pages can eventually lead to the broad pages that manage to cycle without hitting the philosophy page. Furthermore, there can also be pages with no links that function as dead-end pages. Wikipedia recently underwent an effort to remove all true dead-end pages (pages with zero links).(Wikipedia 2023a) Despite these efforts, there remain pages with no links as far as this study is concerned. For example, many sports pages have a lot of links, but they all lie within tables which are not included in this phenomena. For example, on 2011-12 Exeter City F.C. season, there are tons of links but none in the *content* of the page. All of them are in tables or citations, meaning that this is a dead-end page for the philosophy phenomena. Additionally, this study does not consider links in lists, a choice explained in greater detail in the methods section.

A page's neighbors will remain within semantically related to that page amongst links in the lead. In a study that constructed Wikipedia's network using the first ten links in an article as a node's edges, it was determined that the nodes will form into communities of semantically related terms.(Neven Matas 2015) The mathematics page will be in a community of other topics related to math such as physics. For our sakes, this is an important result as it helps to paint a picture of what the branches stemming from philosophy's neighbors will look like. For example, we can now expect all scientific terms to be connected in communities allowing them all to pass through the science page on their way to the philosophy page.

Beyond some of the quicker results such as the size of the GCC, the average path length to philosophy, the number of disconnected components, and the nature of networks from other link locations, I will also look at the neighbors of the philosophy node. If the philosophy node is removed from the network, how large is the remaining GCC and what is its largest node? What other large components now exist? I hypothesize that the awareness node and its connecting parts will form the basis of the GCC and that the network will not shrink by more than 10%. However, that if awareness were to be removed as well, the GCC would shrink dramatically as the awareness node serves as a bridge between all scientific topics and all locations-based topics (buildings, monuments, historical figures).

3 Methods

All of my analysis and data collection was done using Python 3.10.12.

3.0.1 Finding First Links

get_first_link(page_url)

By far the most difficult task was writing the function that would find the first (or second) link on a wikipedia page. What is an incredibly easy task for the human eye proved to be quite difficult to program. If the task was to get the first *linked content* on a page that would be quite easy. However, the more literal phenomena occurs with the first link to another wikipedia page in the content section of the page that is not in parenthesis and not a citation. To find this took a lot of trial and error as Wikipedia pages vary far more than you might think.

I first tried to use the Wikipedia API. Its *links* attribute would have seemed to be an easy way to grab the links on a page. However, there is no functionality to get the links in order of appearance; instead, they are returned alphabetically. I briefly investigate ways to figure out which of these links came first by parsing the HTML but quickly found that it would just be easier to do the entire thing using the HTML.

To read the wikipedia pages, I used the Requests and Beautiful Soup libraries. I then found all of the paragraph content on the Wikipedia page. I decided to excluded list components (bullet points) from my analysis as I felt they did not meet the same criteria as a link within a paragraph. This means that pages like History of the Administrative Divisions of China or 1965 Palanca Awards will have 'no links' as their are no links in their primary content. In future analysis, I hope to include these links and compare the results to see which is a better measure.

From there, I found all of the hyperlinks in each paragraph, then got the href and class for each link. I used these to filter out any "bad" links such as citations, files, links that leave Wikipedia, and the most challenging, links within text parenthesis. This was a difficult decision as sometimes it would seem that the text here is meaningful. For example, the Creativity

page has six parenthetical links before you reach the first link. However, most links inside parentheses are for translations and other self-referring content as you can see on the Ancient Greece page; referring to Greek here is wrong as that is in reference to the translation, not the content of the page itself. Therefore, parenthetical links were excluded using the *isValid(ref, paragraph)* function from Christopher Chiche on Stack Overflow.

I then built a list of links and grabbed the first one. Typically, these href's were structured as /wiki/href. To get cleaner node names, I removed the /wiki/ as it would be repetitive to see it as a prefix on every single page. If there were no functional links on a page, it served as a dead-end for the network even though it may not be under Wikipedia's definition of a Dead-End page. This effectively only applied for Disambiguation Pages. If an *AttributeError* or *TypeError* occurs, which is rare, a unique string "!FAIL!:" is added to the front of the url to be detected later so as not to be confused with successful pages.

Finally, to find the second link on a page, I used a nearly identical function that returned the second item in the list or, if there was only one link, no links at all.

3.0.2 Creating the First-Link Network

network_expander(G, page_url, seen_pages, is_root, fails, disconnects,convergence_df, new_pages=100)

This function is used to create or expand the network. This is done using a **Breadth-First Search (BFS)**. It takes in a lot of variables but many of those are just set as empty lists. It is primarily there to give the option of expanding the network in multiple steps rather than one giant run-through as it takes quite some time to run.

The function first checks if there have been any previous iterations or if it is starting new. It also has a list of "Notable Nodes" that I have manually set. These are the nodes that I have found in my analysis to be the most important (central) and therefore want to track their centrality to ensure the network has converged so that we can make claims on the centrality of these nodes despite not encompassing all of the pages of Wikipedia. Additionally, the function monitors the average page distance from the philosophy page and the size of the network's weakly connected Giant Connected Component (GCC) as I will explain shortly. Finally, the function finds the name of the first page it will look at by splitting its url.

Then, the bulk of the function occurs in a for loop. Each time through the loop adds a new "seed page" to the network. Meaning, it starts at a new page and works its way towards the philosophy page or, to another page that loops back to itself. The *new_pages* parameter determines how many times this loop runs. For my final network, I set this to 50,000. However, this does not mean there are 50,000 pages in the network. Rather, there are 50,000 pages plus all of the pages in between those seed pages and the philosophy page, resulting in **INSERT FINAL NUMBER HERE** pages. With greater time and computing power, I would like to

conduct a larger analysis, however, all of the values I discuss would not change in any significant way as demonstrated in the convergence section of my analysis.

Each loop starts with the url of its seed page (*page_url*). For all but the first page, these pages are found using *wiki_random_page(seen_pages)*. This function uses a while loop that ends when the function finds a new random page. It knows it is new if it is not in the input parameter, *seen_pages*, which contains a list of every previous page that the function has seen. It then uses the Wikipedia API's random function to select a random wikipedia page. The function then checks that it has not seen that page before. Then, it avoids two types of pages:

1. **List Pages**: These pages often do not contain any actual information and are just lists of other Wikipedia pages. While some would work for the network, many are unnecessary and lack any links in their primary content, creating issues for the network. See List of painters by name beginning with "P" as an example. These are not pages that would impact Wikipedia's navigability and therefore we can exclude them as seed pages. They are not skipped if they are the first link on a page which can occur (e.g. Sitting).
2. **Disambiguation Pages**: These pages were a much easier decision to skip as they contain no information. They serve to point users to actual pages when their search term was too vague. Additionally, they all lack a first link and would skew statistics such as the size of the GCC. See Category: Disambiugation Pages for more information and the Art Disambiguation Page as an example.

Finally, *wiki_random_page* creates a proper page url by replace spaces with underscores and breaks the while loop. The function then returns the *random_page* name and its url, *page_url*.

It then gets the first link on that page using *get_first_link(page_url=page_url)*. It then double checks that *get_first_link* returned a string. If it did not, and returned a *NoneType* instead, there are two options: 1. If that page is a seed page, it picks a new seed page and starts over. 2. If it is the first link of a different seed page, its url is added to the *fails* list which is manually checked at the end to repair any issues. These are, however, very rare (about 1 in 10,000).

Next, it normalizes the formatting for the first link by making it all lowercase. This is stored as a separate variable as capitalization is [case sensitive in Wikipedia Urls](https://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style/Linking#:~:text=(Wikipedia%20article%20titles%20almost%20always,characters%20after%20the%20initial%20one.). For the network however, the capital p Philosophy page should be no different than the lowercase p philosophy page. Therefore, all nodes are lowercase.

Then, it checks for if the unique fail string, "!FAIL!:" as discussed above in the string. If it is there, it follows the same procedure as if the function returned a *NoneType*, but instead adds the first link to the fails list.

If there is a dead end node, that is added to a list of disconnects, then a new seed page is found using *wiki_random_page(seen_pages)*.

If it makes it through all of those checks, which most pages do, it is added as a node. Additionally, if it is not the seed page, an edge is added from the previous page to its first link.

Then, once all of the Notable Nodes are in the network (which should happen after just a few iterations), at every 1/100th of the total network size, centrality measures of Betweeness Centrality, Closeness Centrality, and In-Degree Centrality are calculated for each of the notable nodes using NetworkX Centrality functions as well as the average distance from the philosophy page and the size of the GCC which are calculated manually. These are then organized into a row of a Pandas DataFrame called *convergence_df*. Out-Degree centrality was excluded because the out-degree of every page is 1, making all of their Out-Degree centralities identical. Additionally, eigenvector centrality was ignored as the idea that high degree nodes would be connected to one another is doubtful in this network. There is nothing to suggest that having a common first link makes that page itself also common. Given that the underlying assumption behind eigenvector centrality is not met by this network, it was not worth tracking. These are the values we will track to ensure the DataFrame is large enough that these values are no longer changing. This is by far the most time consuming part of the function. Once the network reaches a large enough size, these calculations can take several minutes, hence why they are only done 1% of the time to maximize efficiency.

Finally, the first link is checked to determine if it is the philosophy page. In which case, we can now move on to a new seed page as we know its outcome. Then, it is checked to see if it returned to the seed page, meaning it looped back to itself; these are added to the list of disconnects. Most disconnects, however, are found in more central pages and have to be found later looking at Smaller Connected Components. They are then checked to see if they have already been visited, in which case we know their eventual outcome and can select a new seed page. Finally, if none of these conditions are met, it is added to the seen pages and searched for its own first link. This process continues until one of the previous criteria is met.

After the loop is completed, it returns the Network (G), its *seen_pages*, *fails*, *disconnects*, and the Convergence DataFrame (*convergence_df*) to be analyzed.

Incidentally, the function can process roughly 1000 seed pages every 10 minutes, however, this slows down as the network expands due to the convergence calculations. Hence, the network size is limited.

I then manually check and fix any failed pages to complete the network and it is saved to a gml and the convergence data to a csv file.

3.0.3 Creating the Second-Link Network

second_link_network_expander(G, page_url, seen_pages, is_root, fails, disconnects, convergence_df, new_pages=100)

The second link network was created in a near identical fashion with a couple of key differences. First, it uses *get_second_link(page_url)* to gather the pages second links, rather than their first. Additionally, since it is more difficult to know the "notable" pages, these are selected by finding the top 3 pages in each centrality measure. Since we want to wait until a critical size to look into this, it does not start until the network has at least 3000 seed pages. There is also no reason to check each pages distance from the philosophy page in this network so that convergence measure has been eliminated. Similarly, it no longer stops looking for new pages at a specific page like it did for the philosophy page; now, it continues until it hits a page that we have already visited.

I chose not to investigate any other link locations for a few reasons. First, as the link location grows, it becomes increasingly unlikely that a page has that many links on it, shrinking the network. Second, no patterns presented themselves within the second-link network that seemed necessary to investigate at other locations. And finally, as previously discussed, the opening line(s) of a Wikipedia article have signicantly more guidance onto their structure and links (Wikipedia 2023c). Link locations beyond these are unlikely to be more than a random assortment of pages with no notable patterns, however, work would have to be done to prove this claim.

3.0.4 Plotting Methods

To create the plots needed for my analysis, I used MatPlotLib, Seaborn, and NetworkX's Drawing Tool.

4 Results

4.1 First Link Network

The first link network finished with **INSERT NODE TOTAL HERE**.

4.1.1 Convergence

First, it is important to demonstrate that the size of the network is sufficient to make claims as to the nature of this phenomena. Beginning with the centrality measures of the most important nodes in the network, we can see that they have all leveled off and that any additional nodes would only not change their values in any statistically significant way.

In the left column, you can see the centrality measure across the entire network construction while the right column features the last iterations of the network construction to give a "zoomed in" view of the measures. All of them are almost entirely flat, with slopes that are less than 0.001.

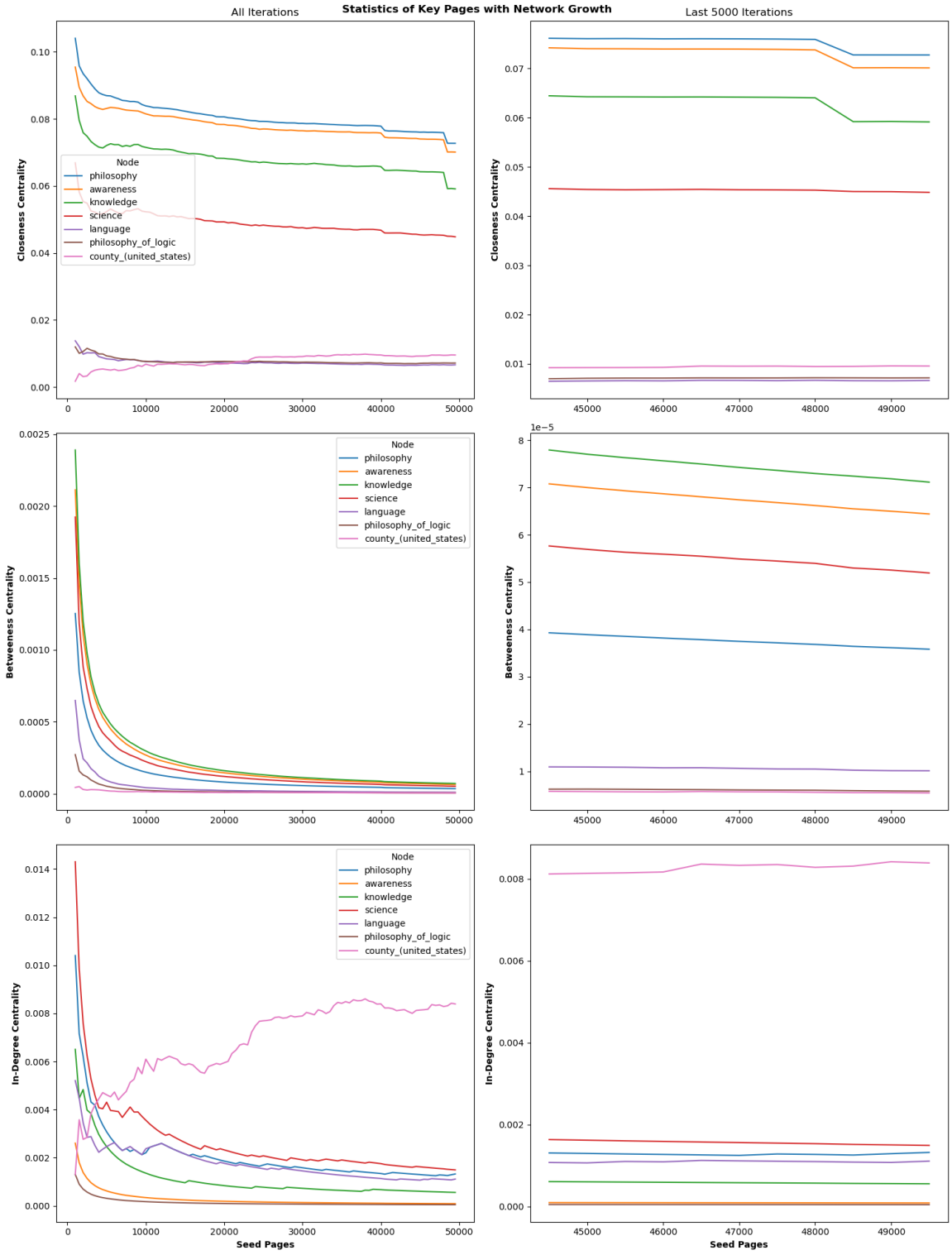


Figure 1: “Node Centrality Convergence”

Next, we can see that the average distance away from the philosophy page also flattened out with the networks expansion.

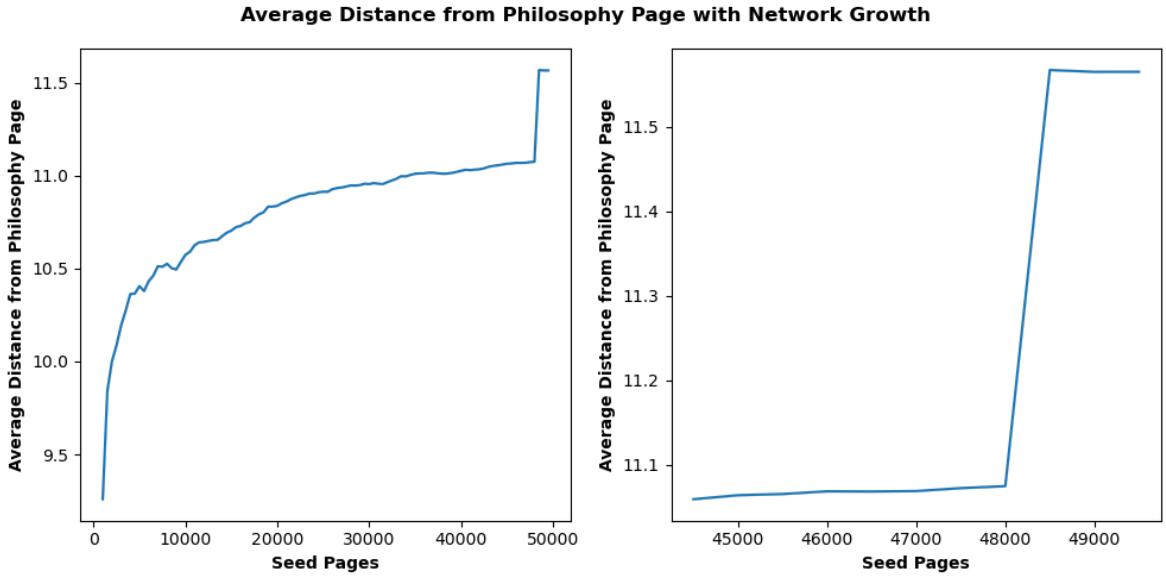


Figure 2: "Average Distance from the Philosophy Page with Network Expansion"

While it is not quite as flat as the convergences, we can still see strong evidence that it has settled to an approximate value of **INSERT HERE**.

Finally, we see an interesting negative slope in the size of the GCC of the network as new nodes are added.

We see that about **INSERT HERE** percent of pages end up at the philosophy page.

4.1.2 Notable Nodes and Paths to Philosophy

As was discussed earlier, there were several nodes that became apparent as the most important in the network. These nodes 'funneled' pages into the philosophy page and thus boasted the largest centrality measures in the network. This can perhaps best be scene by visualizing the nodes closest to the philosophy node using the force-directed Kamada-Kawai Layout.

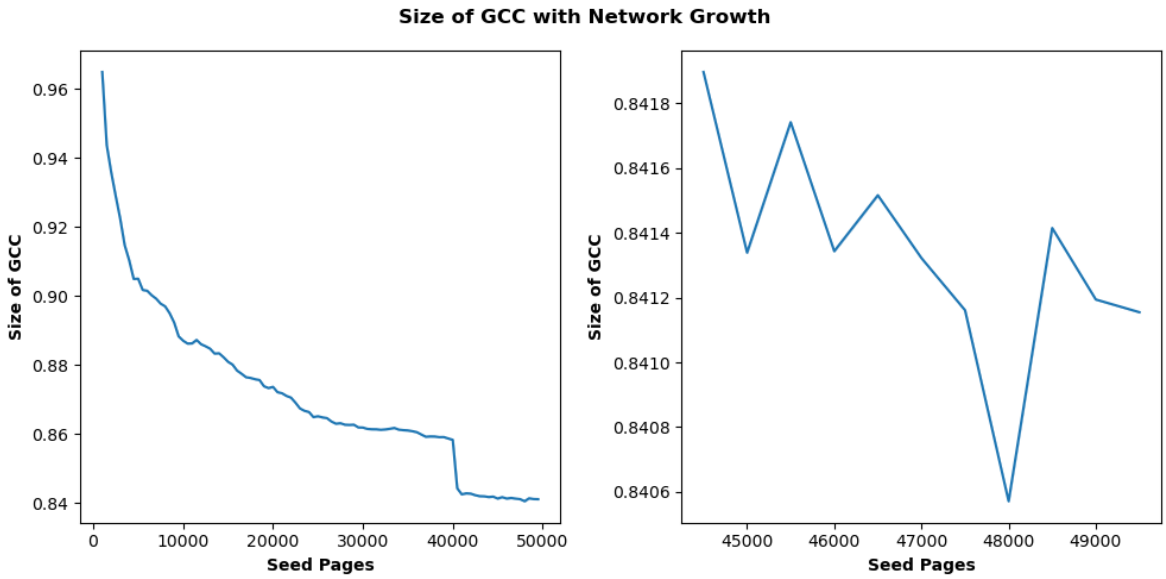
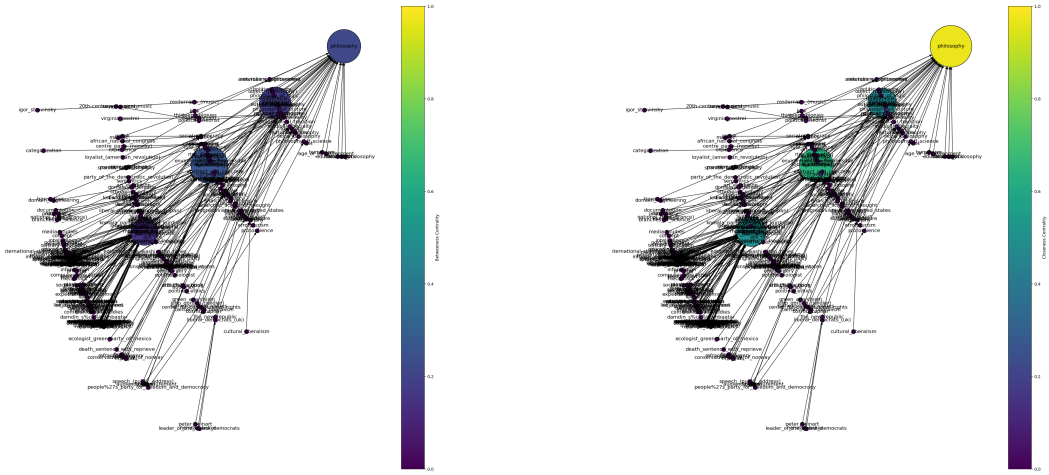


Figure 3: "Size of the GCC with Network Expansion"



These plots help visualize how nodes flow towards the philosophy page. You can see that it has several low degree neighbors in addition to some of these huge hubs. Within my search, I found 31 neighbors of the philosophy page but there are surely countless others that were not found within my search. For example, Immanuel Kant's page's first link is philosophy but it is unlikely to be the first link on more than a handful of other pages, making its chances of being found in a network of this size incredibly small. The same likely goes for numerous

philosophers and adjacent topics. The neighbors of Philosophy are listed below by degree:

	Node	Degree
1	political_philosophy	15
2	specialty_(medicine)	7
3	modernism	6
4	aesthetics	6
5	medical_specialty	5
6	philosophy_of_culture	3
7	ethics	3
8	awareness	3
9	aesthetic	3
10	platonism	2
11	object_(philosophy)	2
12	post-structuralist	2
13	philosophies	2
14	medical_speciality	2
15	american_enlightenment	2
16	political_theory	2
17	natural_philosophy	2
18	moral_philosophy	2
19	naturalism_(philosophy)	2
20	philosophy_of_logic	2
21	philosophy_of_science	2
22	outline_of_philosophy	2
23	subjectivity	2
24	philosophical_tradition	2
25	metaphysics	2
26	humanism	1
27	educational_philosophy	1
28	deist	1
29	neuroethics	1
30	rationalism	1
31	age_of_enlightenment	1

However, this order does not represent the average path to philosophy. More typically, pages will end up in one of a few specific paths to philosophy. For most disciplines in arts, sciences, or technology, they will most often end up on the science page, leading to the knowledge page, then awareness, before philosophy. Awareness is by far the largest neighbor of Philosophy. Its closeness centrality just barely tails philosophy for the second largest in the network. Those two are followed by knowledge and science, respectively. Below are the full values by closeness centrality:

	Node	Closeness Centrality
1	philosophy	0.0727122
2	awareness	0.0700715
3	knowledge	0.0590686
4	science	0.0447865
5	geography	0.0254912
6	continent	0.0226938
7	mind	0.0218818
8	state_(polity)	0.0211012
9	psychology	0.0206636
10	thought	0.0196981

To better understand how nodes reach the philosophy page, here are the top ten nodes by appearances in paths to philosophy as well as the percentage of paths they appear in:

INSERT TABLE HERE

The Awareness node is so central, in fact, that when you remove the Philosophy node from the network, severing Awareness from all of the other paths to philosophy, the GCC is still **INSERT PERCENTAGE HERE** of the network, centered on the awareness page. The next largest component of this network is headed by the Philosophy of Logic page. This is likely because many technical, particularly foreign origin words, will go to their language of origins page. These are then directed to the language page which eventually reaches the Philosophy of Logic page, which then hits the philosophy page. Unfortunately, these nodes are too far from the Philosophy page to be visualized. Furthermore, both the communication and information pages also take you to the Philosophy of Logic page.

There are also several pages with high in-degree centralities that are not seen in these plots. Below are the largest nodes by Degree and In-Degree centrality and their paths to philosophy:

	Node	De-gree	In-Degree Cen-trality	Path to Philosophy
1	county_(united_states)	200	0.0084372	[‘county_(united_states)’, ‘united_states’, ‘north_america’, ‘continent’, ‘geography’, ‘science’, ‘knowledge’, ‘awareness’, ‘philosophy’]
2	associa-tion_foot-ball	158	0.0066300	[‘association_football’, ‘team_sport’, ‘sport’, ‘physical_activity’, ‘exercise’, ‘human_body’, ‘human’, ‘species’, ‘biology’, ‘science’, ‘knowledge’, ‘awareness’, ‘philosophy’]

	Node	Degree	In-Degree Centrality	Path to Philosophy
3	public_university	128	0.0053631	[‘public_university’, ‘university’, ‘educational_institution’, ‘education’, ‘knowledge’, ‘awareness’, ‘philosophy’]
4	u.s._state	109	0.0045608	[‘u.s._state’, ‘united_states’, ‘north_america’, ‘continent’, ‘geography’, ‘science’, ‘knowledge’, ‘awareness’, ‘philosophy’]
5	family_(biology)	94	0.0039273	[‘family_(biology)’, ‘taxonomic_rank’, ‘biology’, ‘science’, ‘knowledge’, ‘awareness’, ‘philosophy’]
6	county_seat	78	0.0032516	[‘county_seat’, ‘seat_of_government’, ‘government’, ‘state_(polity)’, ‘politics’, ‘decision-making’, ‘psychology’, ‘mind’, ‘thought’, ‘consciousness’, ‘awareness’, ‘philosophy’]
7	tennis	74	0.0030827	[‘tennis’, ‘list_of_racket_sports’, ‘game’, ‘play_(activity)’, ‘recreational’, ‘leisure’, ‘time’, ‘sequence’, ‘mathematics’, ‘knowledge’, ‘awareness’, ‘philosophy’]
8	rock_music	69	0.0028716	[‘rock_music’, ‘genre_(music)’, ‘music’, ‘the_arts’, ‘creativity’, ‘psychology’, ‘mind’, ‘thought’, ‘consciousness’, ‘awareness’, ‘philosophy’]
9	capital_city	69	0.0028716	[‘capital_city’, ‘municipality’, ‘administrative_division’, ‘sovereign_state’, ‘state_(polity)’, ‘politics’, ‘decision-making’, ‘psychology’, ‘mind’, ‘thought’, ‘consciousness’, ‘awareness’, ‘philosophy’]
10	administrative_division	65	0.0027027	[‘administrative_division’, ‘sovereign_state’, ‘state_(polity)’, ‘politics’, ‘decision-making’, ‘psychology’, ‘mind’, ‘thought’, ‘consciousness’, ‘awareness’, ‘philosophy’]

As you can see, many relate to geography. The outlier here seems to be the Association Football page which appears here due to a shockingly large number of football (soccer) pages. Whether this is due to random chance in the search or if football pages make up a large portion of Wikipedia's network is impossible to say due to our sample size, however, *association_football* showed up consistently as one of the largest nodes by in-degree during the data collection process.

Finally, the furthest page from philosophy, that reached it, was **INSERT HERE** and its path can be seen below.

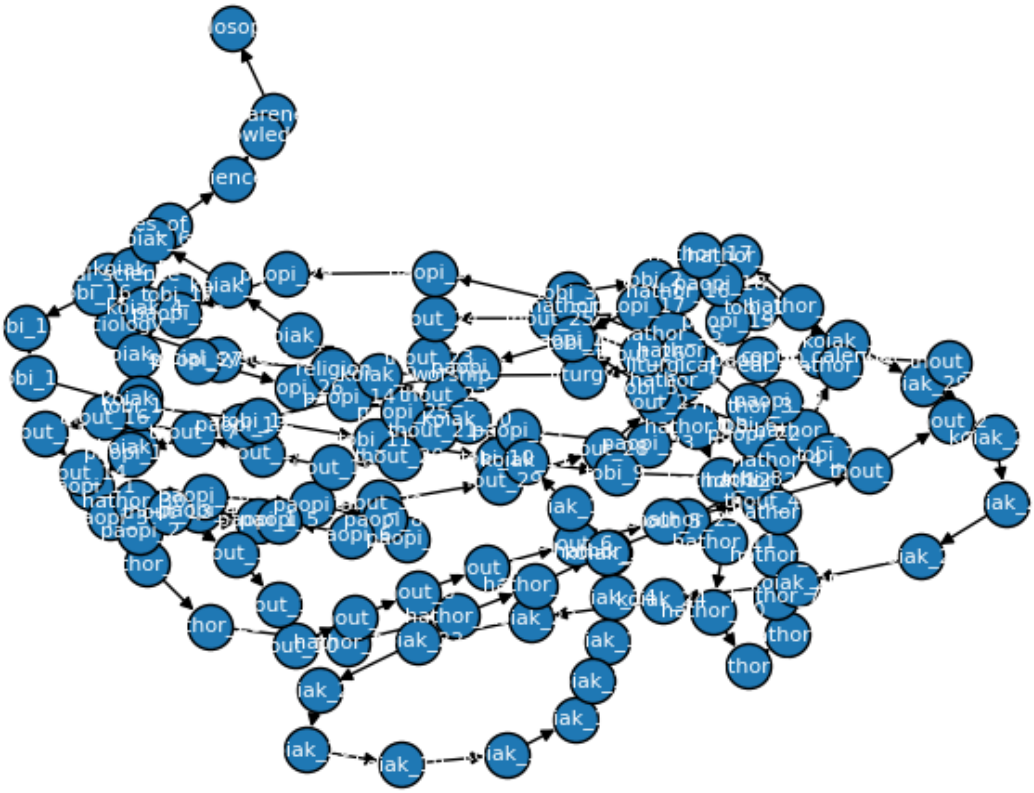


Figure 4: “Longest Path to Philosophy”

4.1.3 First-Link Network Structure

The network also exhibits a typical long-tail distribution. Most nodes have an in-degree of 1, with only a tiny fraction exceeding double-digits.

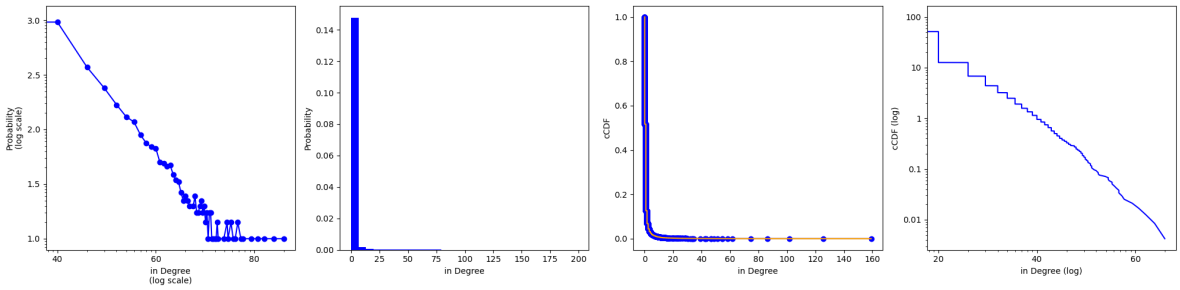


Figure 5: "First-Link Network Structure"

4.1.4 Degree vs. Distance from the Philosophy Page

I also plotted degree vs the distance from the philosophy page to see if there was any correlation,

However, as you can see in the figure above, there was little to no obvious correlation due to the long tail distribution of the network.

4.2 Second Link Network

The second-link network finished with 14,631 total nodes.

4.2.1 Second Link Convergence

Again, you can see that the network's largest nodes all converge, showing little change in their centrality measures during the final expansions of the network.

Similarly, we can see the GCC of the network converges to a little over 16%. While not as flat as I would like, I feel confident that these values would remain around 16% as the network expanded.

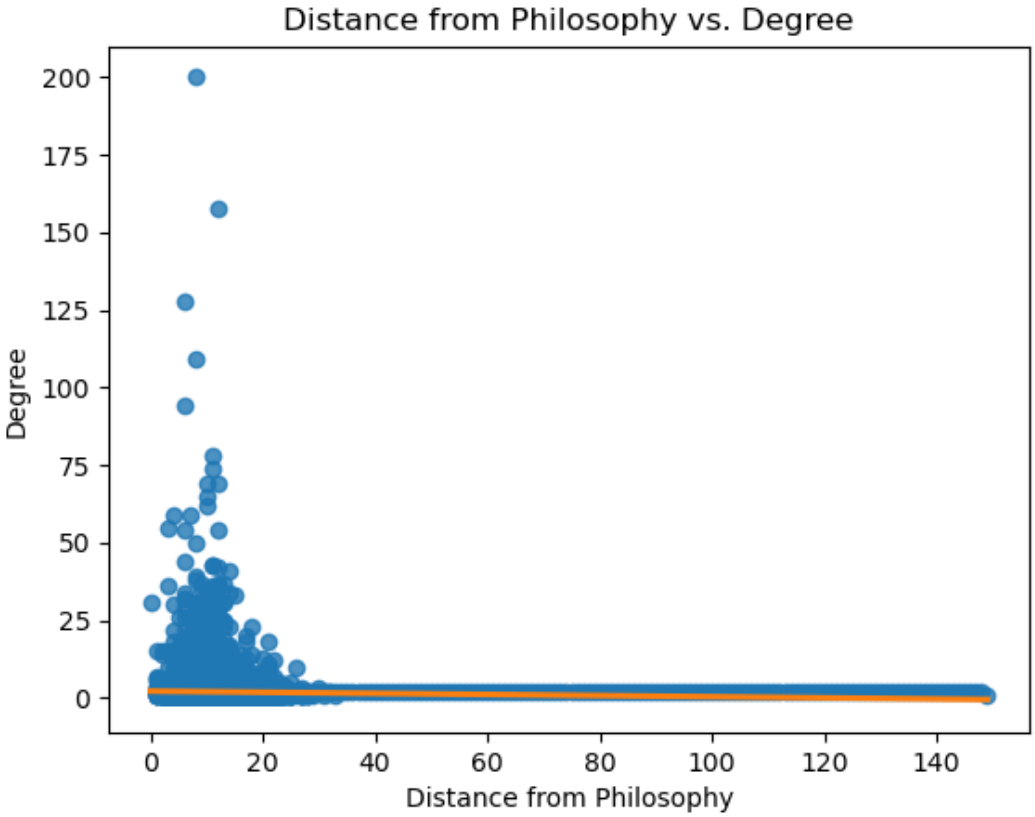


Figure 6: "Degree vs. Distance from the Philosophy Page"

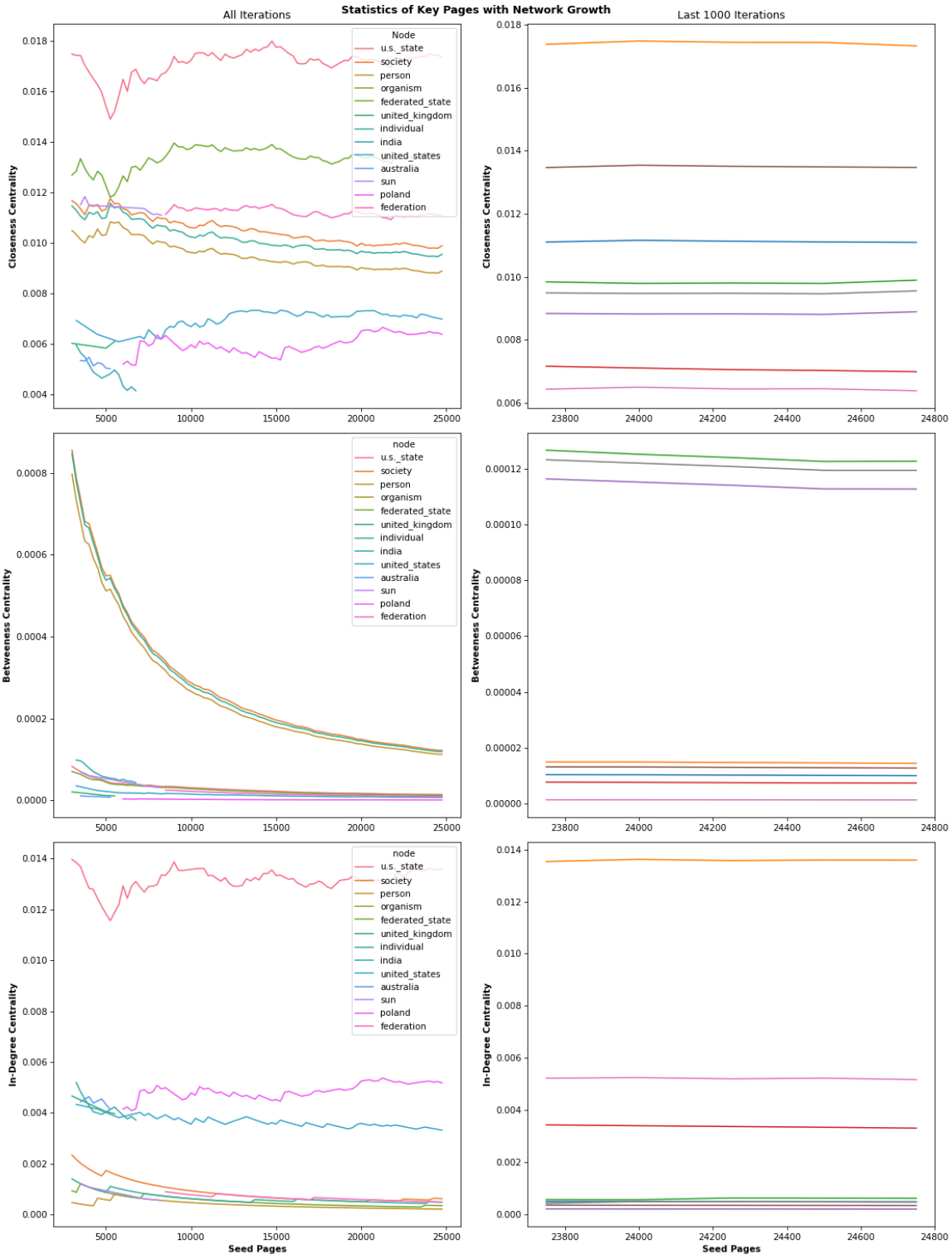


Figure 7: "Second Link Node Centrality Convergence"

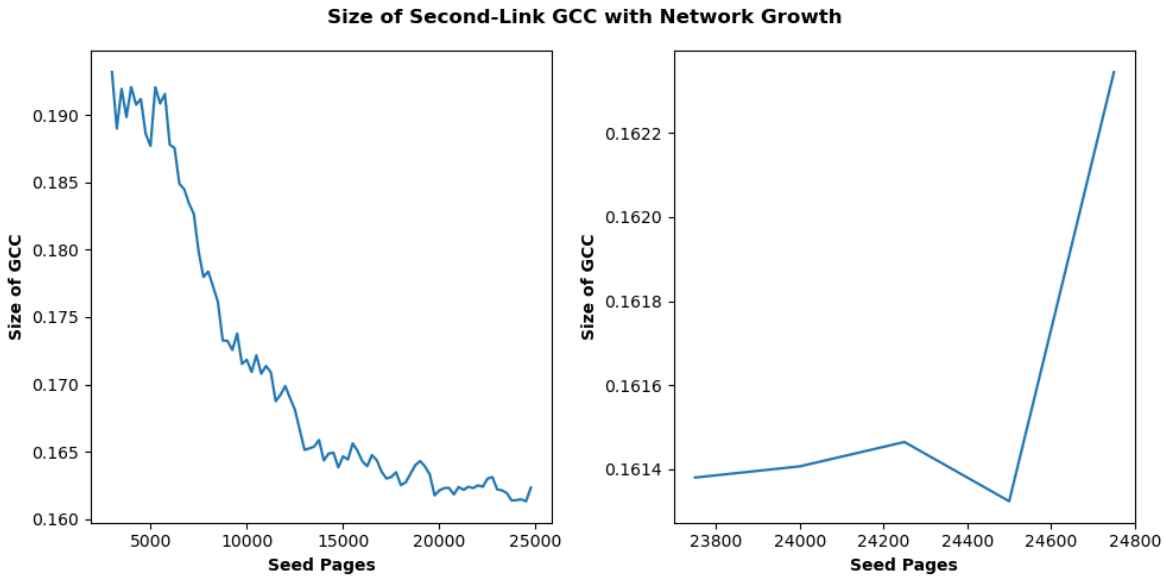


Figure 8: "Second Link GCC Convergence"

4.2.2 Second Link Largest Connected Components

More importantly, we can see here that the largest connected component of the second link network is quite small, at just around 16%. This component's largest nodes are primarily geographic, led by India and Japan. It is visualized in the plot below.

The next largest component is significantly smaller, at just 7.74% of the network. It is similarly led by geographic pages; its top 5 pages were France, Europe, Germany, Spain, and Earth. It is visualized below:

4.2.3 Second Link Notable Nodes

As it may be becoming clear, geographic nodes are the most prevalent in the second link network. By far the largest node by degree, was the US State page. Its peers were almost all geographic pages as well. Here is the full list by degree:

	Node	Degree	In-Degree Centrality
1	u.s._state	198	0.0134655
2	poland	77	0.00519481
3	united_states	50	0.00334928
4	india	33	0.00218729
5	association_football	32	0.00211893

	Node	Degree	In-Degree Centrality
6	united_kingdom	30	0.00198223
7	australia	28	0.00184552
8	research_university	28	0.00184552
9	iran	27	0.00177717
10	france	27	0.00177717

The top nodes by closeness centrality were also led by the US State page, but it was then followed up by more similar terms relating to federalism and governance. Here are the top ten nodes by Closeness Centrality:

	Node	Closeness Centrality
1	u.s._state	0.0172532
2	federated_state	0.013413
3	federation	0.0110436
4	political_union	0.0105248
5	society	0.00991123
6	individual	0.00956221
7	organism	0.00955794
8	sun	0.00938739
9	administrative_subdivision	0.00932515
10	person	0.0089184

Conversely, the most central nodes by betweeness centrality were more philosophic terms. Here are the top ten:

	Node	Betweeness Centrality
1	society	0.000114381
2	individual	0.000110876
3	person	0.000104335
4	organism	0.000101765
5	living_system	9.78309e-05
6	morality	9.69432e-05
7	self-organization	9.2766e-05
8	social_actions	8.95421e-05
9	social_science	8.85655e-05
10	action_(philosophy)	8.244e-05

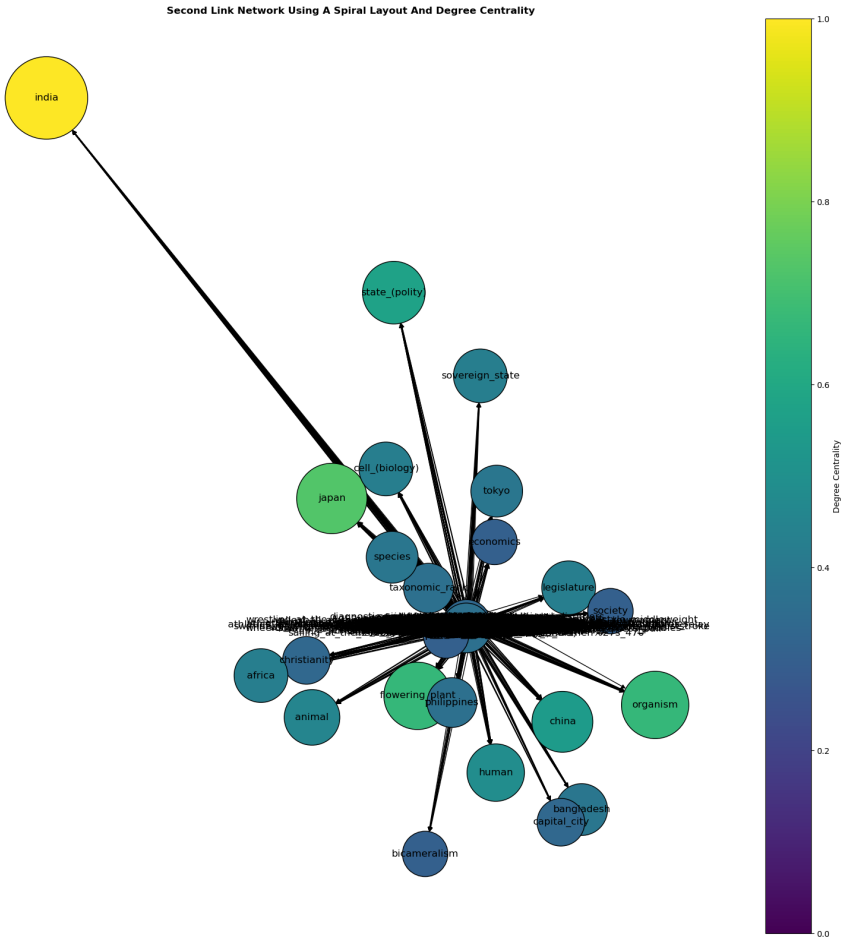


Figure 9: “Second-Link Network GCC”

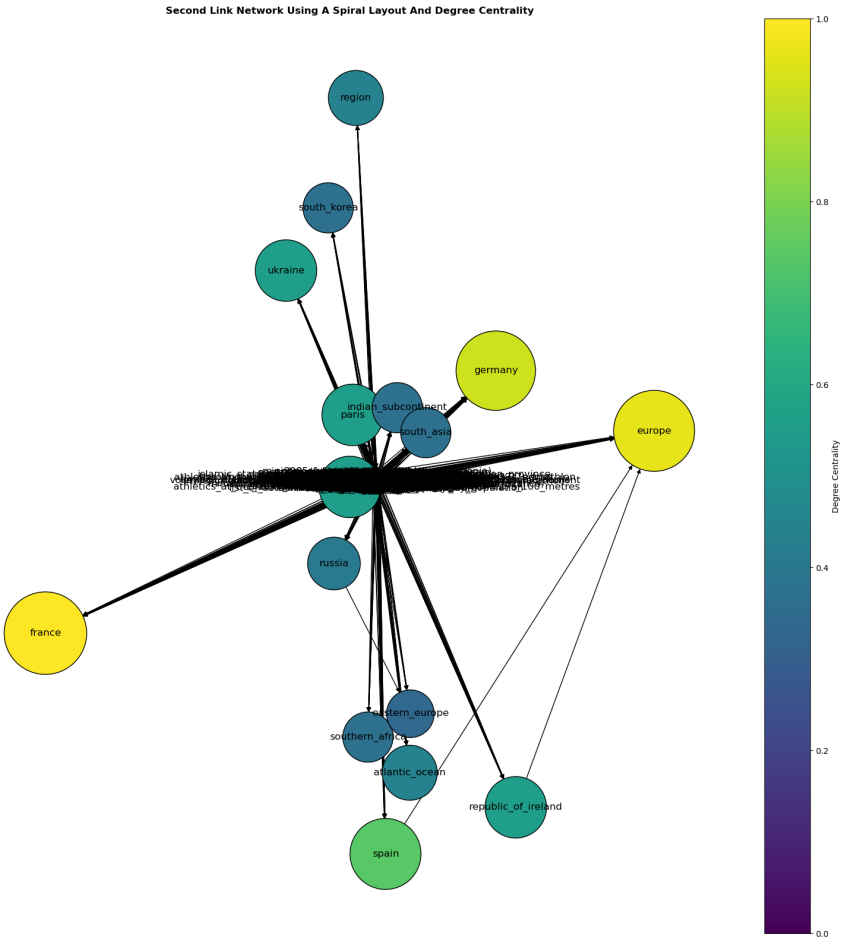


Figure 10: “Second-Link Network Second Largest Connected Component”

4.2.4 Second Link Network Structure

Finally, the second link network also displayed a long tail distribution, with the majority of nodes having a degree of <10 as seen below.

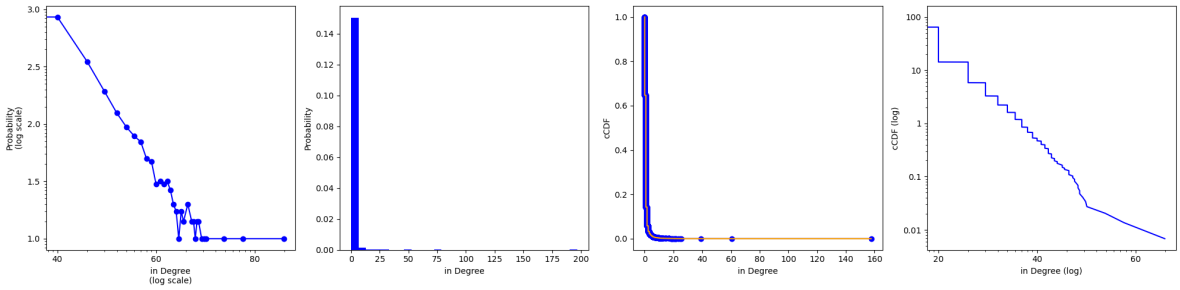


Figure 11: "Second Link Degree Distribution"

5 Conclusions

5.1 First-Link Network Conclusions

5.1.1 Getting to Philosophy

First and foremost, it is clear that the Getting to Philosophy phenomena still holds true. With the vast majority of pages successfully leading to the philosophy page as you click on the first link of the page and every following page. While there was little doubt about this, it is good to see it confirmed.

Within this, we found that it takes, on average, **INSERT AVERAGE HERE** to reach the philosophy page. However, you may find that it does not feel that way when you investigate it for yourself. Since we were looking at totally random Wikipedia pages, many of the pages were very specific. Unless you are quite imaginative, your tests may seem to be much closer to the philosophy page than the average node is.

Additionally, our analysis suggests the network had shrunk from previous research's measurements. It will take further research to understand why this negative slope exists, however, we see it settle around **FIX THIS 87% of the total network**. This is a sharp drop from the previous value of 97% (Daniel Lamprecht 2016). There are a couple of reasons for this. First, I would expect that they included lists in their analysis. More importantly, the 97% figure cited by Wikipedia is not actually the number of pages that reach the Philosophy page from the first link network. Rather, it is "the percentage of articles which eventually lead to a cycle when repeatedly following first links." (Daniel Lamprecht 2016) There are several of these cycles which occur without ever connecting to the Philosophy page. For example, the first link on

the money page is payment. Then, the first link on the payment page is money. This loop blocks several nodes from ever reaching the philosophy page. Similar loops occur on the name, accounting, and candidate pages to name a few. This difference in methodology likely makes up for the near difference 10% difference here. Lamprecht et al. do list a figure for the amount of pages that link directly to the philosophy page of 92.1%. This much smaller difference is likely the result of including list links, the size of our networks, their use of the Wikipedia API, and changes in Wikipedia's network structure over the past seven years. Due to too many interacting variables, it is difficult to make a strong claim as to the true difference in the size of this component.

5.1.2 Why is the Philosophy Node Significant and What is the Nature of the Most Common First Links?

One of the primary questions going into this study was why this phenomena occurs. After all, there seems no obvious answer on first thought. My philosophy professors would tell you it is because philosophy is "the first science" and is the study of nearly everything by that regard. Others might speculate that it is because it is, by nature, a meta discipline, contemplating the nature of other sciences. For any of these reasons and more, it appears as the first link on a few key pages. Namely, by being the first link on the Awareness and the Philosophy of Logic pages. By being the first link on just these two pages alone, Philosophy is already connected to **INSERT HERE** percent of the network.

Then, in order to not lead to an even more central node, philosophy has to link back to itself. The first link network is so connected in the first-place due to Wikipedia's instructions to make the first link on a page increasingly broad(Wikipedia 2023c). This ensures that few pages are capable of these loops. We found that the most common first links often dealt with geography. This means that Wikipedia's instructions, as laid out in the introduction, were executed quite well. Pages are explained in their broadest terms first, particularly, by locating the topic in the world. This results in county, state, and other geographic terms taking the lead. Why *association_football* is so common remains a mystery, the prevalence of football (soccer) pages on Wikipedia would be an interesting subject to explore in the future.

Philosophy follows a path back to itself that is actually quite a bit longer than many of the cycles found in the network, many of which bounce right back to themselves from their first link. Philosophy, on the other hand, takes five additional pages (existence, entity, abstraction, rules, philosophy of logic) to get back to itself. Thus, technically, all of these pages are connected to such a vast portion of the network. These pages, are merely riding on philosophy's coattails in this technicality. Philosophy remains a dramatically more central node. Therefore, no matter why you think the neighbors of philosophy are so connected to so many pages, philosophy's connection to just a few important nodes and its self-loop are the root of this phenomena.

5.1.3 Distance from Philosophy and Degree

Here, I had hypothesized to see that being closer to the philosophy page meant your topic was broader, resulting in a higher degree. This would have been quite an interesting discovery, but, unfortunately, I found no significance. Whether such a significance could arise with a greater network size is unclear, but I remain curious as to whether this may be. I expect that the more specific topics near the philosophy page, such as specific philosophers or topics, drags this down. Additionally, my analysis was quite simple, with more time I would like to look into a more complex statistical analysis here to see if there is a correlation that extends beyond the eye.

5.2 Second-Link Network Conclusions

There is not as much to say here. First, no page presented itself as exhibiting anything similar to what was occurring with the philosophy page in the first-link network. I am further convinced that this is a phenomena unique to the first link location on Wikipedia pages. If the second link, which is the least random link besides the first, cannot form a connected component larger than even just 20% of the network, one must imagine that any other link location would form increasingly random and disconnected networks.

The second-link network was able to create nice loops due to similar reasons as discussed with the first network, their links are prescribed to be broad. Again, you see geography take an even greater role in this network, likely because it still exists for buildings where the larger geographic context (such as first link city, second link state) is still present. Meanwhile, many objects that are described first go straight to their geographical location instead of the nature of what they are. Then, these locations would get back to themselves by looping on their broader locational significance, as a state in a federalist system, for example.

While interesting in some regards, this network serves more as a reason to dismiss further studies of other link locations as they are increasingly unlikely to demonstrate a pattern of similar note to the first link network.

5.3 References

Daniel Lamprecht, Markus Strohmaier. 2016. "Evaluating and Improving Navigability of Wikipedia: A Comparative Study of Eight Language Editions." *OpenSym '16: Proceedings of the 12th International Symposium on Open Collaboration.*

Dimitar Dimitrov, Markus Stromaier. 2016. "Visual Positions of Links and Clicks on Wikipedia." *WWW '16 Companion: Proceedings of the 25th International Conference Companion on World Wide Web* 2.

Neven Matas, Ana Meštrović. 2015. "Extracting Domain Knowledge by Complex Networks Analysis of Wikipedia Entries." *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)* 3.

Wikipedia. 2023a. "Wikipedia:dead-End_pages." 2023. https://en.wikipedia.org/wiki/Wikipedia:Dead-end_pages.

———. 2023b. "Wikipedia:getting to Philosophy." 2023. https://en.wikipedia.org/wiki/Wikipedia:Getting_to_Philosophy.

———. 2023c. "Wikipedia:manual of Style." 2023. https://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style.

xefer. 2011. "All Roads Lead to "Philosophy"." 2011. https://www.xefer.com/2011/05/wikipedia.