What's So Special About Philosophy?

Unraveling Wikipedia's First Link Network

Abstract

Apples, oranges, and the most obscure Dylan song too—is any article a few clicks from Philosophy? The surprising answer is yes, 95% of the time (according to a popular blog post).

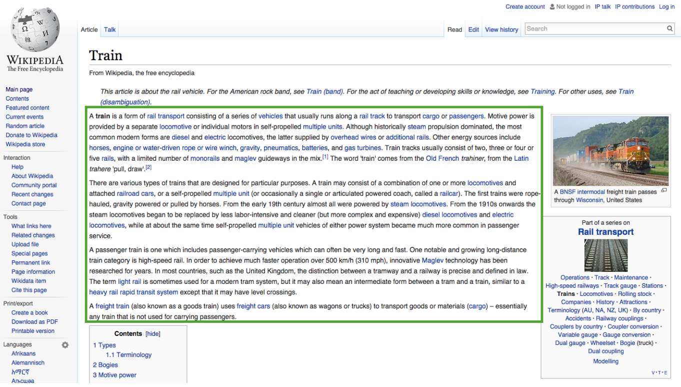
Wikipedia is the largest, most meticulously indexed collection of human knowledge ever amassed. Yet, we have never closely examined the curious web tying one entry to another. By following the first link in an article, we connect entries to form a directed network within Wikipedia: Wikpedia's First Link Network. Our aim is to study Wikipedia's First Link Network for insight into the topics, ideas, people, objects, and events we link. ((improve this sentence))

We first map the First Link Network by algorithmically parsing the first link. Then, apply a river network model to identify the basins of attraction, loop structure, and characterize the many paths from one page's first link to the next. ((wording is unclear: as if there are many paths from one page))

rewrite abstract

At no point in history has a larger or more meticulously indexed collection of human knowledge existed [1]. In amassing such an awe-inspiring collection, we naturally linked topics, inventions, people, objects, places, and events [1] across space and time. The web we weaved, and continue to weave, is a wealth of information about those notable inventions, places, figures, and ideas, but also about us: the ones weaving the web. Here we explore this web as a directed network formed by the first link connecting one article to another.

We build Wikipedia's First link network by following the first link, not in parenthesis, inside the main body of each article of the English version of Wikipedia. For the directed network to meaningfully reflect how we associate one article to another, we exclude links in parenthesis containing pronunciation keys or disambiguations. We also exclude any links in the side-bar elements, as well as any links to external pages, files, or other WikiMedia projects outstide of Wikipedia such as Wiktionary ((lookup others)) [2] thus, forming a true network of Wikipedia.



Train

From Wikipedia, the free encyclopedia

This article is about the rail vehicle. For the American rock band, see Train (band). For the act of teaching or developing skills or knowledge, see Training. For other uses, see Train (disambiguation).

A **train** is a form of rail transport consisting of a series of vehicles that usually runs along a rail track to transport cargo or passengers. Motive power is provided by a separate locomotive or individual motors in self-propelled multiple units. Although historically steam propulsion dominated, the most common modern forms are diesel and electric locomotives, the latter supplied by overhead wires or additional rails. Other energy sources include horses, engine or water-driven rope or wire winch, gravity, pneumatics, batteries, and gas turbines. Train tracks usually consist of two, three or four or five rails, with a limited number of monorails and maglev guideways in the mix.[1] The word 'train' comes from the Old French *trahiner*, from the Latin *trahere* 'pull, draw'.[2]

There are various types of trains that are designed for particular purposes. A train may consist of a combination of one or more locomotives and attached railroad cars, or a self-propelled multiple unit (or occasionally a single or articulated powered coach, called a railcar). The first trains were rope-hauled, gravity powered or pulled by horses. From the early 19th century almost all were powered by steam locomotives. From the 1910s onwards the steam locomotives began to be replaced by less labor-intensive and cleaner (but more complex and expensive) diesel locomotives and electric locomotives, while at about the same time self-propelled multiple unit vehicles of either power system became much more common in passenger service.

A passenger train is one which includes passenger-carrying vehicles which can often be very long and fast. One notable and growing long-distance train category is high-speed rail. In order to achieve much faster operation over 500 km/h (310 mph), innovative Maglev technology has been researched for years. In most countries, such as the United Kingdom, the distinction between a tramway and a railway is precise and defined in law. The term light rail is sometimes used for a modern tram system, but it may also mean an intermediate form between a tram and a train, similar to a heavy rail rapid transit system except that it may have level crossings.

A freight train (also known as a goods train) uses freight cars (also known as wagons or trucks) to transport goods or materials (cargo) – essentially any train that is not used for carrying passengers.

Contents [hide]
1 Types
1.1 Terminology
2 Bogies
3 Motive power

A BNSF intermodal freight train passes through Wisconsin, United States

Part of a series on
Rail transport

Operations · Track · Maintenance · High-speed railways · Track gauge · Stations · **Trains** · Locomotives · Rolling stock · Companies · History · Attractions · Terminology (AU, NA, NZ, UK) · By country · Accidents · Railway couplings · Couplers by country · Coupler conversion · Variable gauge · Gauge conversion · Dual gauge · Wheelset · Bogie (truck) · Dual coupling
Modelling

V · T · E

[1] while anyone with internet access is free to contribute "neutral, cited information," it's worth noting the population of contributors is not necessarily representative of the overall english speaking population. Notably, the contributors are overwhelmingly male ((cite, expand)).

[2] These constraints are the same which led to the popular blog post/s claiming nearly every article is within fewer than 20 first links away from philosophy.

I. Making a Map: Finding the First Links

To map Wikipedia's First Link Network, we use the freely-available xml dump of all 4.7 million English articles [2]. We analyze the snapshot provided on November 2014, representing the state of Wikipedia at the time. Knowing Wikipedia is an ever-evolving project with 10 edits every second and 750 new article per day on average [3], our aim is to characterize the dynamics of the First Link Network, not to record a particular link between one page to another.

The xml dump provided demarkates article elements in in Media Wiki markup, a markup language with special syntax for links. Media Wiki markup also includes templates for audio files, images, side-bar information. While a human can accurately idenfity the first link, to map the entire First Link Network of 4.7 million pages, we programmatically untangled the body text from side-bar, header box, and bad link elements.

While some libraries exist for Media Wiki Markup, we opted to develop our own algorithm for parsing the first link in the xml version of each article. [3] Our parsing algorithm aimed to: 1. squash the initial bugs 2. eliminate the need for several passes. To process an article with one pass, we developed a hierachical system of flags:

- Flag 1: inside template?
  - Flag 2: inside <ref>, <div>?
    - Flag 3: inside ()?

The algorithm loops in three characters chunks, shifting by one character steps through the article markup. Proceeding in three character chuncks eliminates any nesting issues. If any flag triggers are detected the Flag is raised. Once a flag is raised, we stop processing and proceed to the next character. A first link is correctly identified only if Flags 1, 2, and 3 are all off. Then the entire link is retrieved. We then confirm whether the link is legitimate, meaning it does not link to an image, audio file, external page or other Wikimedia project via keyword triggers provided by Wiki Media [4]. The first link of an article is then the earliest link in the text to pass all three flags and the false link check.

Instead of relying on a sample of articles from which to generalize, we opted to process all 11.blah articles in the xml dump containing the 4.7 million articles along with redirected pages and disambugations, which we filter. To process the entirety of Wikipedia, we divided the xml dump into collections of articles, distributing the processing across 112 cores of the UVM sueprcomputer cluster. We then obtained a complete map of the First Link Network.

testing the algorithm

II. Using the Map: Navigating the First Links

reword the shit below quote analysis from other areas/dodds paper: Droplets accumulating

first paper by Dodds cites Horton-Strahler (leaf pruning) not drop accumulation

To understand the structure of Wikipedia's First Link Network, we need to understand how the links flow from one page to another. The procedure, inspired by work in networks/enviromental something ((ask dodds)) is to treat each page as a starting point.

We follow the path of the first links until we reach either a dead link–a page linking to an invalid link (a file, another wikimedia project, or an external webpage)– or a loop, where one page in the path is revisted.

What we can then measure are various metrics including the number of pages whose path leads to a particular page ((confusing wording)), the path length, even loops of a particular size. Combining these measures, we can

[3] Approaches using existing libraries led to several bugs including trouble with nested links, nested parenthesis, unclosed tags, escape characters as well as compatibility with other libraries used to parse the xml.

then examine are particular properties of how pages related: for example the most popular two-loops.

Thus, from the directed network of First Links, we associate to each node several properties. Fortunately, we need not worry about issues of sampling error or statistiacal accuracy, since we do not rely on any sampling. Instead, thanks to the VACC we are able to follow every single possible path, tallying the properties of each node in the network.

Whether philosophy or another page is at the center of Wikipedia's First Link Network is a question about the page with the most traverses. More importantly, instead of a mere page, we can precisely identify the exact loop structure (or basin) where most pages feed.

Not only this, we can answer whether the top loop or page is in fact special–whether the number of pages feeding into the page/loop is measurably different from the other top pages/basins.

((add diagram to explain flow of procedure))

III. The Stories of our Journey: our results

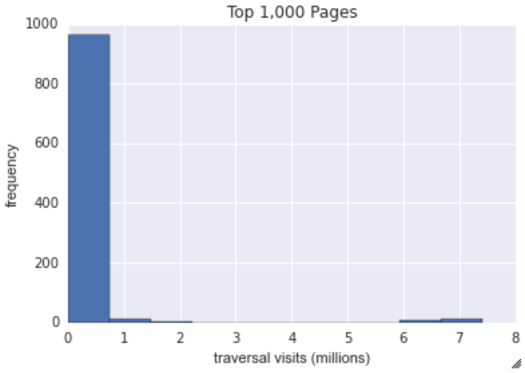
III.1 Clicks (beans...)

Distribution is zipf's law: ((plot)) 8.6 million of the 11 million pages in our raw dataset (11 million includes redirects to the subset of 4.2 million actual pages) have no clicks, with less than 1% of pages accruing more than 100 clicks (precisely 8.56 million of the 11 have no visits).

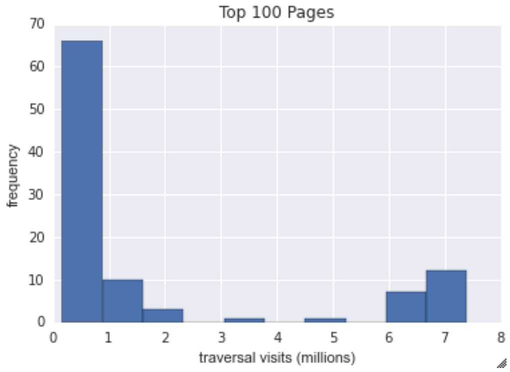
((maybe sample and graph log-log to see if it's linear to justify zipf's law))

average number of visits is 21.

Even examining only the 1000 pages with the most page visist, we find a majority have no visit:

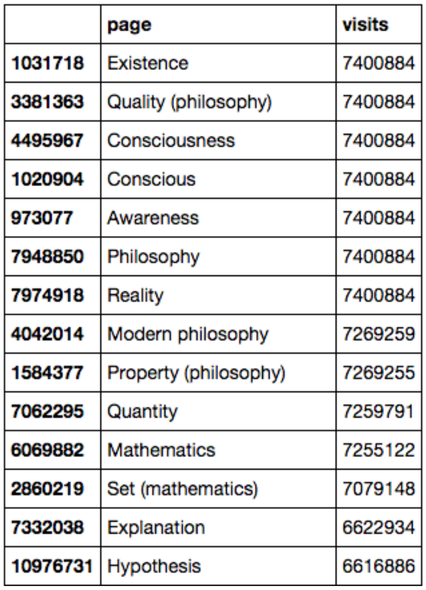


Of the top 100 pages, we see fewer than 30, contain an overwhelming majority of the traversals 6, 7 million:



Maximum number of traversals is 7.4 million, but the mean is a mere 20 visits.

The top 7 pages, the ones with teh most visits comprise a loop around philosophy:



	page	visits
1031718	Existence	7400884
3381363	Quality (philosophy)	7400884
4495967	Consciousness	7400884
1020904	Conscious	7400884
973077	Awareness	7400884
7948850	Philosophy	7400884
7974918	Reality	7400884
4042014	Modern philosophy	7269259
1584377	Property (philosophy)	7269255
7062295	Quantity	7259791
6069882	Mathematics	7255122
2860219	Set (mathematics)	7079148
7332038	Explanation	6622934
10976731	Hypothesis	6616886

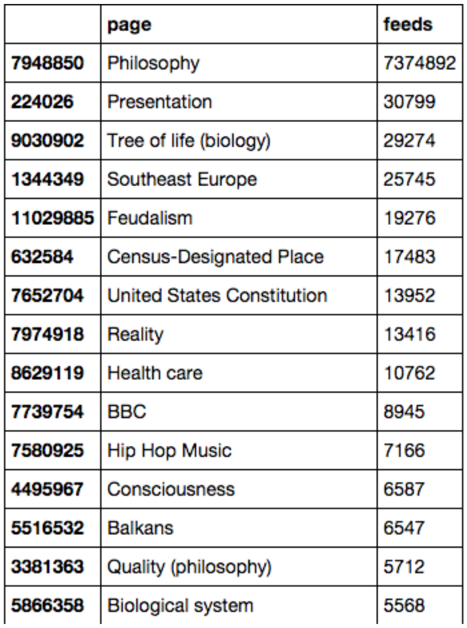
III.2 Longest Path

The longest first link path span around 365 pages. These paths traverse calendar related pages including religious calendars such as Orthodox Liturgics and history pages. The Eastern Orthodox Liturgics pages for example are dated. Visit any page's Orthodox Liturgics and the first link will send to tomorrow's liturgics. The enxt days links to the following day's liturgics. Curiously, these are complete loops. The last calendar day of the year, simply links back to the first restarting the loop. Other notable pages with a path length of 80 include the legislative assembly of Ontario, 1868 in the UK,

III.3 Loops - basics

Although the loop of 7 pages around philosophy does attract the greatest number of visits, several other loops around natural fundamental concepts also exist. The top non-philosophy basins include a basin around community, another around landmass.

Identifying basins explains the structure of the network, but it doesn't reveal which page/s are a funnel for the basin. By following a similar approach and measuring where the drop accumulation is funneled we find:



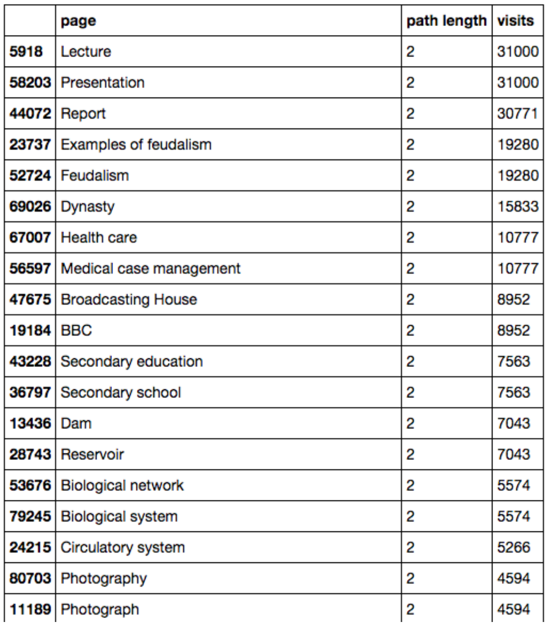
	page	feeds
7948850	Philosophy	7374892
224026	Presentation	30799
9030902	Tree of life (biology)	29274
1344349	Southeast Europe	25745
11029885	Feudalism	19276
632584	Census-Designated Place	17483
7652704	United States Constitution	13952
7974918	Reality	13416
8629119	Health care	10762
7739754	BBC	8945
7580925	Hip Hop Music	7166
4495967	Consciousness	6587
5516532	Balkans	6547
3381363	Quality (philosophy)	5712
5866358	Biological system	5568

confirming philosophy is indeed the largest funnel for the philosophy loop, implying it is in fact at the center of wikipedia. Nonetheless, curious foundational funnels also surface such as the word presentation, Tree of life (biology), Geographical regions such as Southern Europe, the US Constituion, health care and BBC. This amalgamation mixes both foundational topics, confirming teh theory whereby the first link generalizes a specific page by hooking it to a broader concept. The list also highlights broader topics that are perhaps less foundational, but seem to be at the forefront of people's minds. Health Care while not a foundational discipline such as biology or philosophy ranks at the top perhaps due to the importance we as a society attach to it. Similarly BBC, Hip Hop Music surface perhpas out of a cultural ttraction.

how many two loops are there compared to the rest of wikipedia?

In addition to measuring pages that funnel to accumulation, we can also measure true loops instead of just basins. These are pages which sit in a closed loop as opposed ones leading to a dead link.

Furthermore, since we can rank pages, we can rank loops based on the measure associated with each page. We first examine two loops:

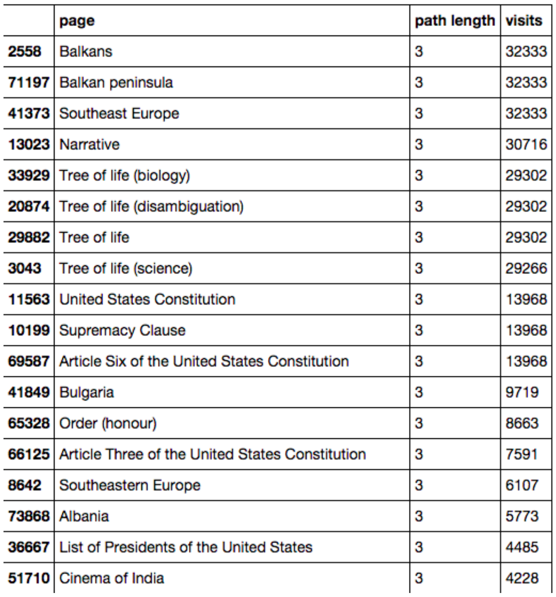


	page	path length	visits
5918	Lecture	2	31000
58203	Presentation	2	31000
44072	Report	2	30771
23737	Examples of feudalism	2	19280
52724	Feudalism	2	19280
69026	Dynasty	2	15833
67007	Health care	2	10777
56597	Medical case management	2	10777
47675	Broadcasting House	2	8952
19184	BBC	2	8952
43228	Secondary education	2	7563
36797	Secondary school	2	7563
13436	Dam	2	7043
28743	Reservoir	2	7043
53676	Biological network	2	5574
79245	Biological system	2	5574
24215	Circulatory system	2	5266
80703	Photography	2	4594
11189	Photograph	2	4594

Two loops reveal ideas that are closely linked to each other or nealry synomous such a secondary school the place wher esecondary education takes place. A lecture which is a form of a presentation, or fedualism and its examples.

Investigating the list beyond simply the top pages, but looking at a typical 2-loop we find a more natural link between (where typical means between 5 and 15 visits) linking a person to what they are most known for: say a book and author: Anatomy of Britain links to Anthony Sampson (and the reverse; an inventor and their invention: Voere and VEC-91; and an event links to the creator of the event: Peotry Bus Tour links to Wave Books.

While often a less intimate connection, 3-loops reveal a similar pattern. The top ranking 3-loops are:



	page	path length	visits
2558	Balkans	3	32333
71197	Balkan peninsula	3	32333
41373	Southeast Europe	3	32333
13023	Narrative	3	30716
33929	Tree of life (biology)	3	29302
20874	Tree of life (disambiguation)	3	29302
29882	Tree of life	3	29302
3043	Tree of life (science)	3	29266
11563	United States Constitution	3	13968
10199	Supremacy Clause	3	13968
69587	Article Six of the United States Constitution	3	13968
41849	Bulgaria	3	9719
65328	Order (honour)	3	8663
66125	Article Three of the United States Constitution	3	7591
8642	Southeastern Europe	3	6107
73868	Albania	3	5773
36667	List of Presidents of the United States	3	4485
51710	Cinema of India	3	4228

There are 74k true 3loops.

III.4 Later: Visualize loops

D3 Example 1

simple D3 Network

IV. Conclusion

Next: word stem to see how broader ideas are related (taxonomy of ideas). * Broader: what if we consider all links (or first par)? other languages? * popular page by links to it.

* Is the network stable over time? * analyze a few newer datasets.

The analysis remarkably demonstrates the links in Wikipedia truly exihibit a power-law like distribution where most pages link up to a handful of pages. Furthermore, Philosophy does occupy not only a special place inside the most popular basin of attraction. Not only that, the philosophy page itself is the page through which the feeds flow. This implies there is something special even from a mathematical standpoint. Philosophy isn't just a popular link, it lies in an exceptionally high percentage, unique to its loop.

Aside from the philosophy loop, other foundational concepts also emerge as the centers for basins of attraction. Fundamental concepts such as community, government, or philosophy. The network also has a prelevance of 2-loops ((find exact proportion or number)) which link two closely associated concepts such as an inventor or the invention as well as tie synonyms together. Other artifacts such as tremendously long loops based around calendar days demonstrate how we group ideas such as holidays, hisotiric events, or religious, not by their ties to other ideas, but first by where they fall on calendar days.

Based on the flow from specific pages to more general concepts, the results also agree with the heuristic provided here. Whereby, we link a very specific page to a broader concept, then to an ever broader concept. For example, the Wikipedia page for Banana links to fruit, then botanical then, ((confirm, fill-in)) then biology, then science and eventually philosophy. It is curious that Wikipedia, the largest encyclopedia, which by its very meaning aims to capture information about all branches of human knowledge, has naturally evolved through the works of ((x million editors/edits)) so that Knowledge, philsophy are the center links, occupying such a special place.

IV.1 Terms (aka style guide)

use article instead of page

traversal = node weight path length = degree of separation from a cluster path is the flow from one page until a dead link or a repeated article

cycle (permutation from abstract algebra) instead of loop

basin of attraction or cluster? as opposed to cycle

invalid link (instead of dead or bad link)

To Do:

1. Add limitations of work: snap shot in time, parsing is imperfect
2. Add better graphics
3. post code and results to github (add link to paper)

References

[1] "The Decline of Wikipedia", MIT Tech Review

[2] Wikipedia Database Download

[3] Wikipedia Statistics

[4] Media Markup Special Pages