

# Connecting Every Bit of Knowledge: Wikipedia's First Link Network Unravalled

Mark Ibrahim,<sup>1,\*</sup> Christopher M. Danforth,<sup>1,†</sup> and Peter Sheridan Dodds<sup>1,‡</sup>

<sup>1</sup>*Department of Mathematics & Statistics, Computational Story Lab,  
The University of Vermont, Burlington, VT 05401.*

Apples, oranges, and the most obscure Dylan song too—is everything a few clicks from Philosophy? Within Wikipedia, the surprising answer is yes: nearly all paths lead to Philosophy. Wikipedia is the largest, most meticulously indexed collection of human knowledge ever amassed. More than information about a topic though, Wikipedia is a marvelous web of naturally emerging relationships. By following the First Link in an article, we connect entries to form a directed network within Wikipedia: Wikipedia's First Link Network. Here we study the English edition of Wikipedia's First Link Network for insight into how we relate topics, ideas, people, objects, and events.

We algorithmically parse all 4.7 million articles to construct a map of Wikipedia's First Link Network. In a novel approach to uncover structure, we traverse every possible path through the network, measuring the accumulation of First Links, path lengths, basins, cycles, and even particular articles funneling links into the cycles. We discover many scale-free distributions, find Philosophy at a salient center, and uncover a flow from specific to general with basins around fundamental notions such as Community, State, and Science. Curiously, we also observe a gravitation towards topical articles including Health Care and Fossil Fuel. These findings enrich our view of how we connect and structure an ever growing load of information.

## I. INTRODUCTION

Wikipedia is a towering achievement of the modern era. At no point in history has a larger or more meticulously indexed collection of human knowledge existed. Wikipedia contains 37 million articles in 283 languages, with coverage spanning everything from little known ancient battles to the latest pharmaceutical drugs [11] [10]. All this information does not lie dormant. Wikipedia is the sixth most visited site in the world, surpassing 18 billion page views in a single month [20]. The most widely used general reference, Wikipedia has amassed an awe-inspiring collection of human knowledge.

It's not surprising then to see Wikipedia as the object of many studies. Researchers have examined the cultural dynamics among editors, [1] the accuracy of the content relative to traditional encyclopedias, [2] [3] the disciplines and topics covered [4] and bias against portions of the population [5]. Wikipedia's content has also proven a powerful tool. Researchers have used Wikipedia to identify missing dictionary entries [6], cluster short text [7], compute semantic relatedness [8], and disambiguate meaning [9].

While these many studies have dissected and fruitfully applied Wikipedia's content, few have examined the connections among the many articles. Beyond information, Wikipedia is a web relating ideas. Within the

body text of Wikipedia articles, hyperlinks reference other Wikipedia articles. A hyperlink from one Wikipedia article to another is a way to relate the ideas conveyed in the two articles [14]. The notion that hyperlinks convey information about the content of a page has proved successful in numerous domains from search engine algorithms such as PageRank [12] to topic classification [13]. We treat a hyperlink as a mechanism to connect two ideas.

The authors of a Wikipedia article choose where and whether to include a hyperlink reference to another Wikipedia article by including hyperlinks in the article's markup. In describing a particular topic, each set of authors decides which articles are the most relevant references. For example, the hyperlinks in the "Train" article depend only on what its authors decide are pertinent articles to reference. The collection of hyperlinks emerges from the independent choice of each article's authors. The first hyperlink to another Wikipedia article marks the earliest moment in the body text where the authors choose to connect the information discussed to that of another article. By following the first hyperlink to another article in the English edition of Wikipedia, we connect the ideas in one article to those in another, ultimately forming a directed network: *Wikipedia's First Link Network* (FLN).

What information could the first links possibly reveal? Inspired by the claim that the majority of first links lead to the "Philosophy" article—popularized by an xkcd comic and subsequently discussed in blog posts [30] [31] [32]—we holistically study Wikipedia's FLN as a map relating areas of human knowledge. The map is a hierarchical taxonomy where "Train" links to a parent

---

\* mark.s.ibrahim@uvm.edu

† chris.danforth@uvm.edu

‡ peter.dodds@uvm.edu

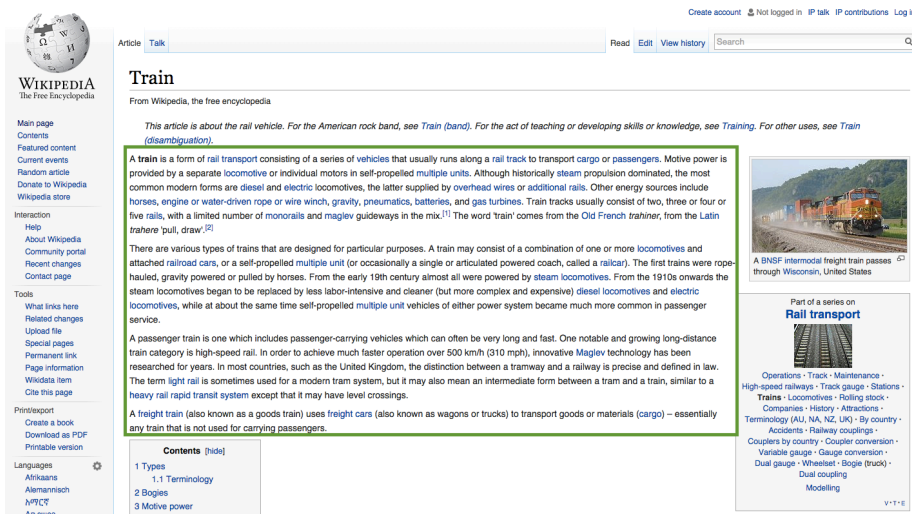


FIG. 1. **First Link Path For "Train"** We follow the first link to another Wikipedia article in the main body of the article—the area inside the green rectangle, which excludes side bar elements, the navigation bar and title. In this example, the first link to another Wikipedia article is "Rail Transport." We can again select the first link on the "Rail Transport" article, repeating the process to form a path of First Links.

node, "Rail Transport," while many child nodes, "Steel" and "Horsepower," link up to "Train." Unlike previous taxonomies created by individuals [15] [17] [16] or a select group of individuals [18], the organization of ideas in the FLN does not emerge from a concerted centralized effort, but rather from the independent choice of every article's authors. The FLN is a wealth of relations among inventions, places, figures, objects, and events across space and time.

Our goal is to gain insight into how the information on Wikipedia is organized and related by studying the structure of the FLN. In a novel approach, we develop metrics to capture the dominant features of the FLN's structure. We measure dynamics of the FLN as a flow, quantifying aspects of the FLN from the accumulation of first links around articles to the influence an article exerts in shaping the FLN. Together with cycles, indegree, depth, and the content of the articles, we build our analysis of the relations among the ideas in Wikipedia.

## II. TRAVERSING THE FIRST LINK NETWORK

One essential measure of a directed network structure is the degree distribution [24]. Degree distribution has been used to study many phenomena from disease outbreak [25] to the dynamics of social networks [26]. The degree distribution in the FLN describes how many first links point to a particular article. We measure the indegree in the FLN to understand which articles authors choose to reference in the first link. A higher indegree indicates more authors selected the particular article as

## First Links "Train"

Rail Transport →  
Conveyance of Passengers and Goods →  
Goods →  
Economics →  
Social Science →  
Academic Disciplines →  
Knowledge →  
Awareness →  
Consciousness →  
Quality →  
Philosophy.

a relevant reference. On the other hand, articles with zero indegree have no references—they are outer leaves in the FLN. Similar to PageRank, the indegree is also a way to rank the articles in Wikipedia. We can rank articles by indegree to see which articles are most and least referenced in the FLN. Additionally, we can assess the degree distribution to see whether authors tend to reference all articles equally or only a few, as observed in many social networks [26].

The degree distribution captures one step in the FLN, measuring the destination of a given link. Indegree measures only the particular set of first links between two articles, rather than the richer dynamics of how many articles fit together—how links create a flow through the FLN. One way to expand the analysis beyond a single step is to map the full path of an article's first link. Each article is both a destination and a link to another article, similar to an element in a linked list data structure. By following the chain of links we can construct a path through the FLN with each article leading to the next via the first link. For example "Train" has a first link to "Conveyance of Passengers and Goods," which is itself an article with a first link to "Goods" and so on. The path starting at "Train" is the collection of articles obtained by following the first links starting at "Train" (see figure 1). The path contains "Goods," "Economics," "Social Science," and other articles related to "Train," but not directly referenced by a first link.

Applying a similar concept with a different starting article, we form a path through the FLN moving from one article to the next. A path in the FLN connects a group of articles allowing us to measure how two arti-

cles relate, even if the articles are not directly linked. The collection of articles is path-connected conveying a flow of concepts from one article to another. The notion of paths allows us to map not only which articles link directly to "Philosophy" for example, but also which ideas are eventually anchored in "Philosophy" even if the articles are not directly linked. The method is order agnostic with respect to which articles are selected first. As long as each article is selected eventually, the resulting metrics are equivalent. Starting at each article, we construct a path through the FLN for each article, ultimately mapping how the first links flow through the FLN.

In a novel approach, we use the notion of a path through the FLN to describe the structure of the FLN as a flow relating ideas. Grounded in the notion of flow through river networks, [27] we characterize features of the FLN as a flow via first links. Previous studies have used flow to characterize the structure of river networks [28] and describe the organization of food systems as transportation networks [29]. With paths to mark flow through the FLN, we develop new metrics to gauge the accumulation of references, the length of the path relating articles, and the influence a particular article exerts in shaping the flow of links through the FLN.

Our goal is to map the flow of first links by following the paths of first links starting at each article. Since a path reveals how the ideas in one article may eventually be anchored in another, the first metric we develop quantifies the accumulation of first links within the FLN. The algorithm begins by selecting an article then traversing the path formed by following the first links. Each time a first link references an article, we increment a count associated with the article. We repeat until the first link is repeated or invalid (outside the FLN). The collection of articles is path-connected conveying a flow of concepts from one article to another. We select a second article and repeat the process until we have constructed a path for each article in the network. We call the resulting count associated with each article the number of *traversal visits*. The number of traversal visits of an article measures the number of references flowing to the ideas in the article.

We can map the paths in the FLN as a matrix with each column corresponding to a path. In our sample network (see figure 2), the path starting at article A is the first column in the traversal visits matrix with 1 if the path contains a given article and 0 if the path doesn't. The second column indicates the path starting at article B and so on. To compute the number traversal visits for article A we sum the corresponding row in our matrix, which totals the number of paths containing article A. The total is analogous to the flow in a river network with traversal visits measuring the accumulation of water near point A. By constructing a traversal visits matrix,

we can similarly compute the number of traversal visits by summing along the corresponding row for our node. The traversal visits matrix for Wikipedia consists 121 million entries to account for each path through the FLN. We measure the accumulation of first link references by summing the corresponding row for the article of interest.

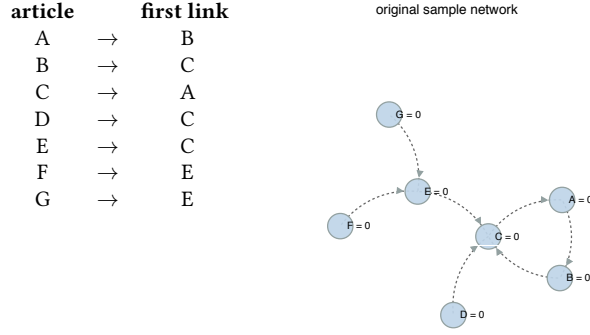
In addition to accumulation, the traversal visits algorithm creates paths which uncover depth in the FLN. By measuring the number of first links between two articles, we get an additional piece of information we call *path length*. We can compute the length a path by summing along a column in our traversal visits matrix. The path length describes how closely related ideas are, gauging the FLN's depth. The greater the number of first links separating two articles, the greater the number of ideas needed to relate the ideas. Although, "Train" is related to "Economics," there are several more specific articles bridging the connection: "Train" is more specifically related to transportation, whose object is often goods. Goods are one of the fundamental objects of study in Economics. Described in links, this relationship is captured by a relatedness of 4 first links ultimately connecting "Train" to "Economics." We can examine path length as a relative measure of how closely connected ideas may be, but also as an aggregate measure describing the overall separation or connectedness in the FLN. We can characterize whether ideas are closely connected by a handful of links, many, or whether a more nuanced organization exists.

One possible path through the FLN is a *cycle* or a group of articles linking to one another inside a loop. In our sample network (see figure 2) a cycle exists among nodes A, B, and C. Since node C links back to the starting node A, the path starting at node A forms a 3-cycle. By recording the history of the articles along a path in the FLN, we can identify the types of cycle structures within the FLN. Furthermore, since each article in the a cycle has an associated traversal visit count, we can rank cycles by the number of references directed towards a cycle. We can also form and rank *basins* in the network by identifying groups of path-connected articles, not necessarily forming a perfect cycle. A basin connects a group of articles which ultimately reference other articles in the FLN. By identifying basins and cycles, we pinpoint group of closely related articles.

While traversal visits measure accumulation, each article's first link also influences the shape of the FLN. At a large point of accumulation, a single article's first link can exert great influence over the shape of the FLN by directing many references towards a particular path. To distinguish between an article that simply happened to fall within a cycle from an article funneling many first links, we develop a second metric called *traversal funnels*.

To measure traversal funnels, we traverse the FLN

## Traversal Algorithm

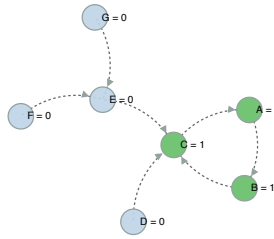


### Traversal Visit Vectors

follow the first link path until a repeated article (or an invalid link)

$$\begin{bmatrix} \vec{A}_{\text{visit}} \\ \vec{B}_{\text{visit}} \\ \vec{C}_{\text{visit}} \\ \vec{D}_{\text{visit}} \\ \vec{E}_{\text{visit}} \\ \vec{F}_{\text{visit}} \\ \vec{G}_{\text{visit}} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

traversal visit path for article A



traversal visit path for article G

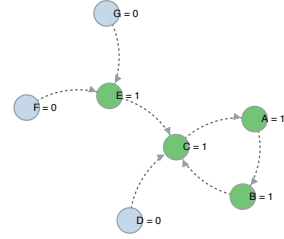


FIG. 2. **Traversal Visit Algorithm on a sample network** The traversal visit vectors are an adjacency matrix for the paths through the network: the first column indicates the path formed starting with article A. The number of traversal visits for article A is then the number of paths containing A or the sum of the first row in our matrix:  $\sum_{i=1}^7 A_{\text{visit}, i} = 7$ .

### Traversal Funnel Vectors

follow the first link path up the start of a cycle (or invalid link)

$$\begin{bmatrix} A_{\text{funnel}} \\ B_{\text{funnel}} \\ C_{\text{funnel}} \\ D_{\text{funnel}} \\ E_{\text{funnel}} \\ F_{\text{funnel}} \\ G_{\text{funnel}} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

traversal funnel path for article G

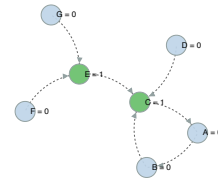


FIG. 3. **Traversal Funnel Algorithm on a sample network** The algorithm for traversal funnels is identical to the previous algorithm for traversal visits with one alteration: the path ends at the start of a cycle to distinguish articles directing a path into a cycle from articles that simply happen to be in a highly traversed path. We can construct similar vectors by considering each path through the network, measuring traversal funnels for a particular article as the sum of the entries in its corresponding row. For example the number of traversal funnels for article E is  $\sum_{i=1}^7 E_{\text{funnel}, i} = 2$ .

in the same manner as we did for traversal visits, but end a path once we enter a cycle. We are then able to distinguish between an article related to many other ideas only by virtue of its place in a cycle from an article

exerting influence over where the first links flow. An article with a large number of traversal funnels directs many references towards a particular path. In our sample network (see figure 3) article C directs the flow of

links towards the 3-cycle, while articles A and B are recipients of the flow—without exerting any influence themselves. While articles A, B, and C all have the same number of traversal visits, article C has 4 traversal funnels (A and B have none). By incrementing the count of first link references on paths only up to a cycle, we can identify which articles exert the greatest influence over the FLN's structure.

By studying the FLN not only as collection of directly linked pairs of articles, but as a flow of paths, we build a powerful arsenal of information with which to study the FLN. From accumulated references, cycles, basins to influence, we can measure how the many articles in Wikipedia are organized and related.

### III. DISCOVERIES

#### A. Degree Distribution

We first identify the set of first links directly referencing a particular article by studying the indegree of the FLN. A higher indegree suggests more authors found the ideas in the article a pertinent direct reference. We rank all 11 million articles by indegree to find the "United States" with 80,249 direct first links as the most referenced Wikipedia article. Other high-ranking articles include foundational abstract concepts such as "village," "species,"; sports associations such as "American Football," "Association Football"; and developed nations such as "France," "Japan," "Russia," "Australia," and the "Netherlands." These high ranking articles are useful abstractions: nations describe a collection of individuals, a common culture, language, or geographical proximity; sports teams describe an ever changing collection of sports players often associated with a cultural identity and a geographical region. Since abstractions such as nations and teams are inherently comprised of many parts, authors reference the larger abstraction when describing a part. For example, in describing the geographical region of north America, the United States, Canada, or Mexico are natural references anchoring the geographical location to a larger national identity.

"Philosophy" and other philosophical concepts with many traversal visits are not among the highest-ranking articles by indegree. "Philosophy" has an indegree of 581, with direct first links from articles about Philosophers and areas of Philosophy: "Existentialism and Humanism," "Predeterminism," "Synoptic Philosophy," "Qualia," "Dorothy Emmet," and "Christopher W. Morris." While many articles accumulate at "Philosophy" (see traversal visits discussion), the accumulation is not the result of many articles directly referencing "Philosophy." Instead, the accumulation of first links, as we argue in our discussion of traversal visits, flows towards Philosophy as the

ultimate anchor when generalizing from specific articles towards broader notions. The FLN's indegree exhibits a scale-free distribution where a few articles receive a most direct first references, while most articles receive few or none. The average indegree for all 11 million articles is 3.6 direct first links with a standard deviation of 89.5 links. Only 4,826 articles have more than 100 direct first links and 75% of articles have fewer than 9. When fit to a linear model on a log-log scale ( $\log(\text{rank})$  versus  $\log(\text{indegree})$ ), the model's  $r$  value is  $-0.98$  suggesting a strong log-log linear fit with a power law exponent of  $-0.79$ .

\*\*\*

Beyond indegree, our work expands the analysis of relations to understand how many more than two articles are organized and related. We develop the notion of a path to describe the structure of the FLN as a flow relating ideas. We traverse every possible path through the FLN, taking 232 million steps along the way.

#### B. Depth of the FLN

We first seek to describe the depth of the FLN: how many links does a connection of ideas span? We gauge the FLN's depth via path length or the number of links traversed until a repeated or invalid link (one outside the FLN). We discover the longest path length is 365 corresponding to the yearly calendar of Orthodox Liturgics. Each day's Liturgics links to the next day's. Curiously, on the last calendar day, the last article simply links back to January 1, forming a 365-cycle (see discussion of cycles). We also found similarly lengthy paths following the evolution of a place or topic through time: "1953 in Scotland" or "1560s Architecture", with articles sequentially proceeding by year, decade or era. The longest paths connecting ideas are organized temporally.

Of the 11 million articles, 5.5 million had an invalid link or linked back to the same article, yielding a path length of zero. This roughly corresponds to the official number of articles on Wikipedia 4.7 million as of November 2014—approximately half of the 11 million articles in the XML dump are redirects or disambiguations, not full articles. The most common path length is 29, with an interquartile range (26, 30). The median path length far exceeds the popular 6-degrees of separation (see for example the average number of academic publications separating scholars [22]), indicating a greater network depth. The network depth suggests a large network of spanning a variety of ideas. As a distribution, more than 75% of articles have a path length below 50 first links while a few temporally organized paths exceed 50 links (see figure 6).

How can we characterize the flow of ideas along a path? Next, we develop a metric to gauge the accu-

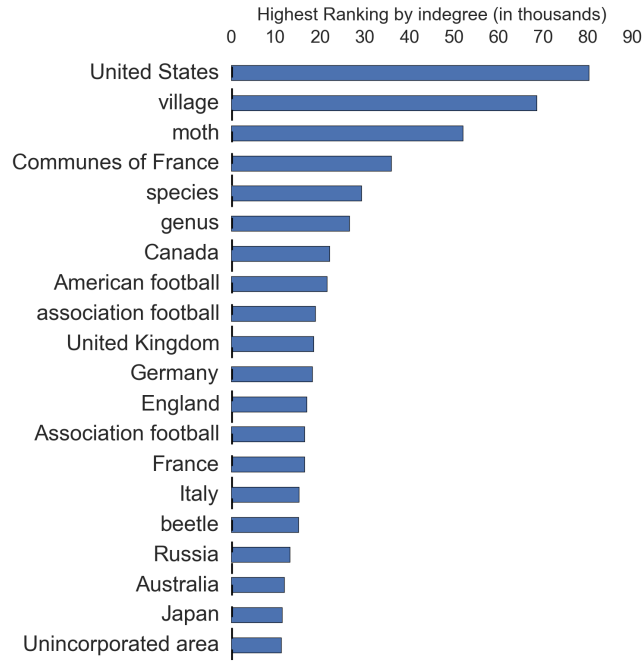


FIG. 4. **Highest Ranking Articles by indegree**

We rank each article by the number of direct first links to the article (indegree). The highest-ranking articles tend to represent abstraction for concepts comprised of various facets and parts.

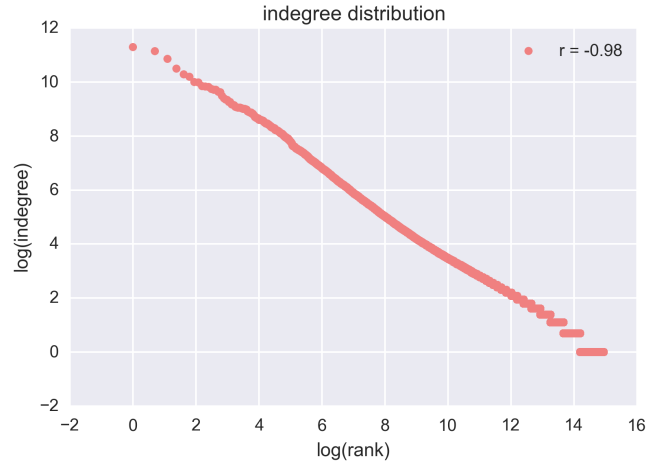


FIG. 5. **FLN Degree Distribution**

We construct a log-log plot fit our results to a linear model. The result is an excellent fit with  $r = -0.98$ , yielding a power law exponent of  $-0.79$ . The distribution appears to be scale free with most articles having fewer than 9 direct first links, while few hold most.

mulation first link references and characterize a global organization of ideas among articles.

### C. Traversal Visits

We followed every possible path through the network, incrementing the number of traversal visits for every

first link reference along a path to generate a total of 232 million traversal visits.

As a distribution, the number traversal visits per article appears to be scale-free. The majority of articles have fewer than 30 traversal visits, while few have 5 order of magnitude more traversal visits. Specifically, 99.76% of articles have fewer than 100 traversal visits; nearly 80% have none. Meanwhile, the highest ranking 30 articles

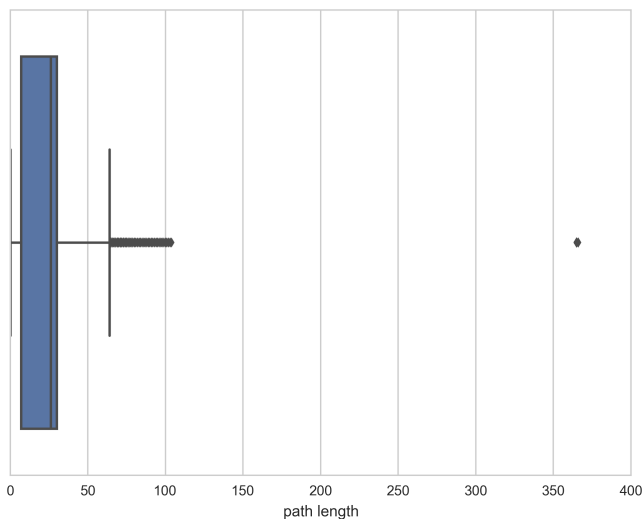


FIG. 6. Path Length Distribution

The median network depth measured by the number of first links in a path is 29. The 365 cycle for "Orthodox" liturgics is the outlier to the right, while other historical articles about "Scotland," "UK" and so on are slightly outside the third quartile. More than 75% of articles have path lengths between 0 and 50 links.

have an extremely disproportionate number of traversal visits.

To more accurately gauge the distribution, we construct a log-log plot of the entire dataset:  $\log(\text{traversal visits})$  against  $\log(\text{rank})$ . We observe a strong linear fit, as is characteristic of scale-free networks, with an  $r$ -value of  $-0.93$  and a power-law exponent of  $-0.636$ . A handful of the highest ranking articles contain a disproportionate number of traversal visits, while most have none. The skew in the distribution is not terribly surprising when considering the heuristic of how the links flow: from specific to general.

The highest ranking articles include Philosophy alongside related articles such as "Existence", "Quality", and "Reality".

#### 1. A flow from general to specific

The highest ranking articles by traversal visits are broad, global topics (see figure 8): many are academic disciplines such as "Science," "Math," "Geography," "Philosophy," "Biology," and "Physics"; others are abstract fundamental concepts such as "Community", "State", "Earth", "Information", "Existence," "Communication", and "Power." Since traversal visits measure the accumulation of First Link visits in Wikipedia, the highest ranking articles suggest a flow from specific to general. For example the flow of First Links for the "Banana" article begins at a very concrete and specific topic then

flows into progressively broader and broader disciplines, eventually culminating at "Philosophy: "Banana" links to the broader category of "Fruit," which then links to "Botany," eventually "Biology," then "Science" and ultimately "Philosophy."

One means to measure the specificity of an article is to identify the number of synonyms available for a word or topic. The reasoning here is that a broader topic likely has many more synonyms relative to a specific, concrete topic. Banana has fewer synonyms than botanical for example. To quantify the observed flow, we measured the number of synonyms each title contained in WordNet—the largest lexical database of the English language [23]. We find the highest ranking 100 articles have on average 5 more synonyms than the typical article; a difference of 2.5 synonyms if we compare the median number of synonyms in each group (see figure 9). As suggested by the median, many articles have no synonyms as we might expect, because titles with more than one word are not likely to appear in a thesaurus. Since many articles have no synonyms, we also compared the number of synonyms in the highest-ranking versus typical article, this time excluding all articles without at least 1 synonym. We still find the highest ranking 100 articles with an average of 9.0 (median of 7.0) synonyms whereas the remaining articles on average 5.8 synonyms (median of 7.0), even with the exclusion of all articles without any synonyms. The quantifiable difference in synonyms corroborates the flow of links from concrete, specific articles to broader disciplines or fundamental notions.

#### D. Network Cycles

The recurrence of an exact number of traversal visits suggests some articles are part of a cycle. The "Philosophy" article for example sits in what seems like a cycle of seven other articles; "Hypothesis" appears to sit in a cycle of 6 other articles including "Experiment", "Fact", and "Knowledge". To confirm the suggested cyclic structure, we record the history of articles traversed along a path. For example starting on "Train" we construct an path of articles: "Rain Transport," "Conveyance of Passengers and Goods," "Goods," "Economics," "Social Science," "Academic Disciplines," "Knowledge," "Awareness," "Consciousness," "Quality," "Philosophy."

We first identified 2-cycles, meaning a pair of articles with First Link pointing to one another. Of the 11 million articles, 84,000 are 2-cycles. The highest ranking 2-cycles by traversal visits tend to be synonyms (or nearly so) rather than distinct, yet connected ideas: "Health Care" and "Medical Case Management", "Broadcasting House" and "BBC", "Secondary Education" and "Secondary School" (see figure 10).



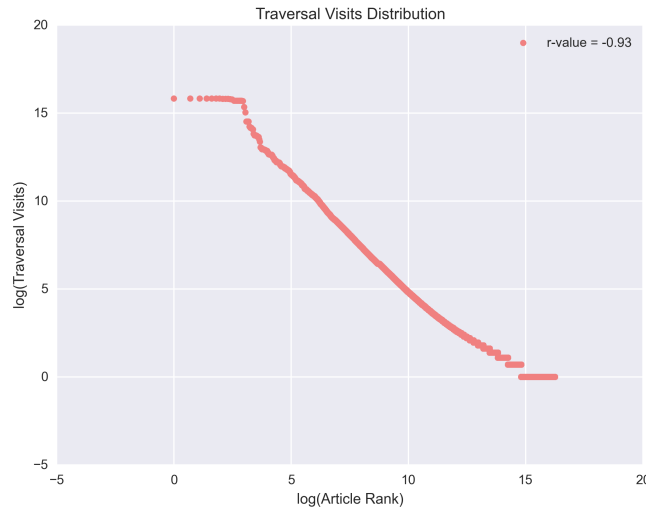


FIG. 7. **Distribution of Traversal Visits**

We fit the distribution to a linear log-log model by considering the log (base 10) transformed rank of each article against log (base 10) transformed the number of traversal visits. The model explains 86% of the variation in the data, yielding an  $r$ -value of  $-0.93$  and a power law exponent of  $-0.636$ . The horizontal flattening around the highest ranking articles is a result of the cyclic structure (see discussion on cycles).

Outside of the highest ranking 2-cycles, the typical 2-cycle signals a connection between distinct, yet very closely related ideas. Link patterns such as inventor to product ("Voere" to "VEC-91"), event to organizer ("Poetry Bus Tour" to "Weave Books"), and book to author ("Anatomy of Britain" to "Anthony Sampson").

Similarly, 3-cycles captured a synonymous or close relation among 3 articles: "Tree of life (Biology)", "Tree of life (disambiguation)", and "Tree of life"; "Cinema of India", "Indian Cinema", and "Telugu Cinema" (see figure 11). Once we extend our cycle size beyond a length of 6 however, "Philosophy" along with the remaining list of high ranking articles by traversal visits dominate.

Extending beyond 3-cycles we find the highest ranking cycles of length up to 10 links are dominated by "Philosophy," high ranking 3-cycles around the "Balkans" and "Tree of life" as well as popular 2-cycles. The longest cycle in the network spans 365 articles of Eastern Orthodox Liturgics for each calendar day. Curiously, on the last calendar day, the last article simply links back to January 1, forming a 365-cycle. Other lengthy cycles span 60-75 articles including collections of articles on national histories such as "Japanese Eras" or judicial bodies such as the "Legislative Assembly of Ontario".

### E. Basins

We can group articles lying on the same path to identify *basins*: a group of path-connected articles. Since cycles identify only groups of articles with a closed set of links, we additionally measure and rank basins to

capture groups of closely related articles branching outside of a cycle into the rest of the FLN. We rank basins by the total number of traversal visits in each of the articles along the path forming the path. Akin to river networks, these basins are areas of accumulation with a path flowing outwards to the rest of the FLN.

The highest ranking basins by the number of traversal visits are groups of articles around "Philosophy." Since the accumulation of first links to "Philosophy" is so high, the paths leading to "Philosophy" carry many references. The highest ranking paths include branches of philosophy flowing through "Awareness," "Existence," and "Consciousness" to "Philosophy." Other paths include concepts around "Mathematics," scientific "Experiments," "Biology," and "Fact." These paths link many specific articles to "Philosophy," each funneled through a particular domain.

Excluding basins around "Philosophy" we find other basins around foundational ideas such as "Community," "Landmass," "Federal Government," "Presentation," and "Belief System." The basins around each of these foundational notions are various paths containing related articles. For example around "Community" we find basins flowing from "United States" to "Federal Republic" to "Political Union" to "State" culminating at "Community"; we also find basins flowing from "Public Policy" to "Executive (government)" to "Government" to "State" and then to "Community"; we also find paths flowing through a similar chain beginning with "Democracy," another beginning with "Constitution," another at "Dictatorship," and so on. The ideas build from specific means of organizing a community (or society) then build



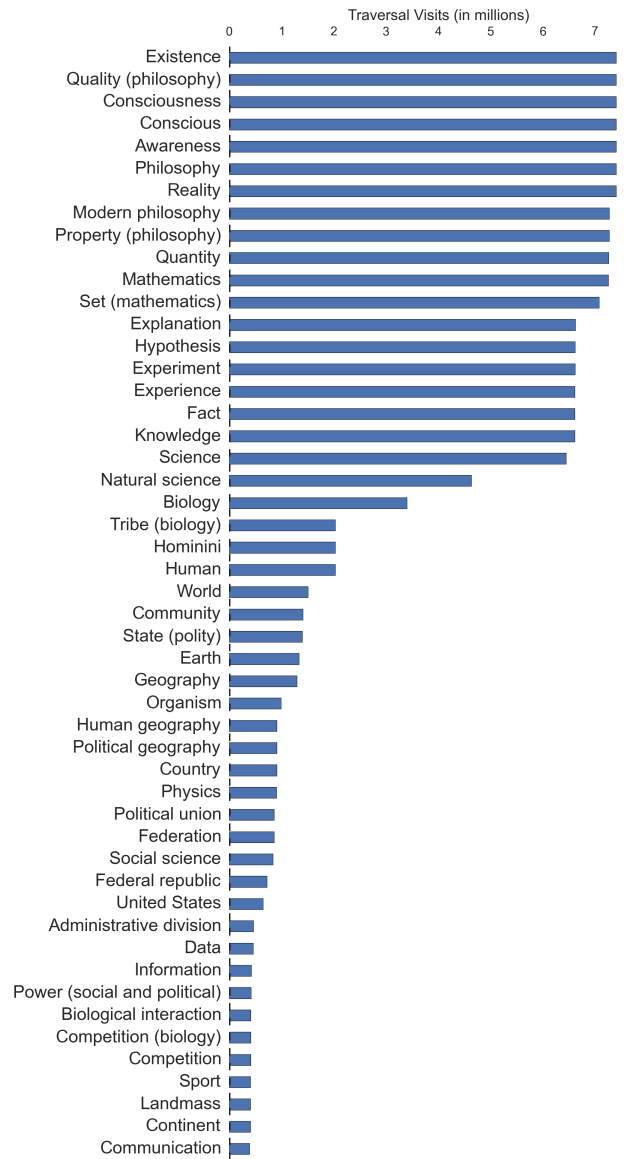


FIG. 8. **highest ranking articles by number of traversal visits**

We compute the number of traversal visits for each article in the FLN (see Traversal Algorithm section for details). In doing so, we can rank each article by the accumulation of first links. Articles with a greater number of traversal visits mark greater points of first link accumulation. The highest ranking articles by traversal visits reveals where the greatest accumulation occurs.

up to "Community." Other basins around landmass for example begin at specific geographical regions such as "Eastern Europe" building up to "Continent" and finally "Landmass." Each basin accumulates first link references through various natural paths linking general themes to the foundational concept.

Which specific articles direct the path towards a particular concepts? Next we study the FLN using traversal funnels to understand the influence a particular article exerts in directing the flow of first link references.

## F. Traversal Funnels

To analyze the influence an article exerts in shaping the structure of the FLN, we compute the number of traversal funnels for each. Articles directing more paths exert a greater influence over the structure of the FLN by increasing the accumulation of first links on a particular path. By measuring traversal funnels, we distinguish between an article that simply happened to fall within a cycle from an article funneling many first links.

Ranking articles by the number of traversal funnels we find "Philosophy" as by far the highest-ranking article

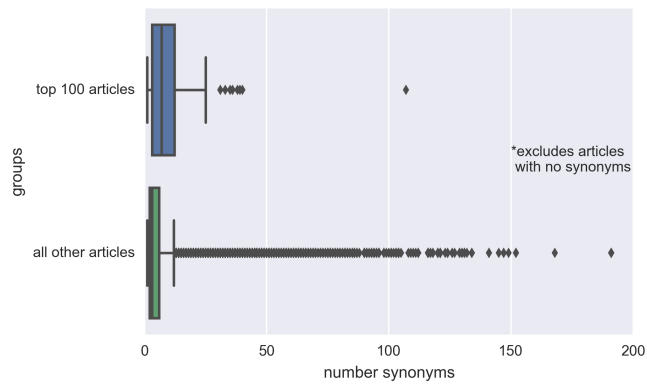


FIG. 9. **Number of Synonyms for the highest-ranking articles** We use WordNet (the largest lexical database for the English language) to identify the number of synonyms for an article's title. We compare the distribution of synonyms of the highest-ranking 100 articles against all other Wikipedia articles. As the distribution demonstrates the highest ranking concepts have more synonyms than other articles.

## 2-cycles

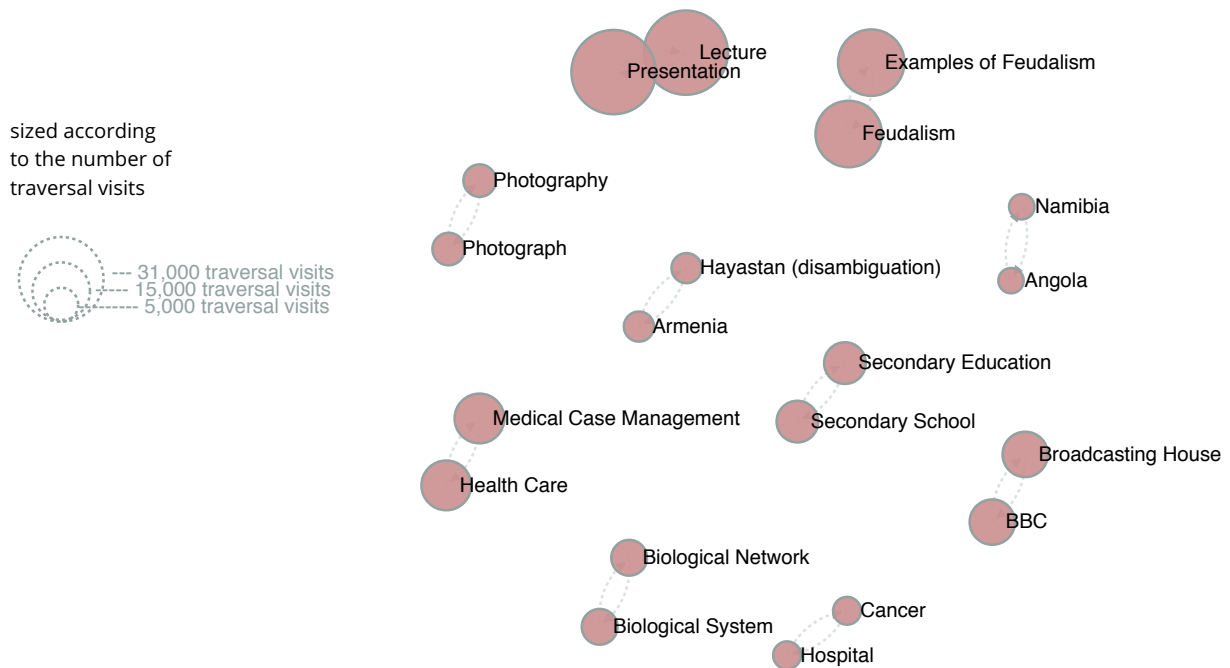


FIG. 10. **highest ranking 2-cycles** We identify pairs of articles whose first links point to one another, forming a 2-cycle. We then rank each pair of articles by the total number of traversal visits to gauge the most referenced groups of two articles linked to each other. We find are synonyms representing the same or nearly the same concepts as opposed to distinct ideas.

with 7.37 million paths (see figure 12). Of any article, the number of traversal funnels Philosophy holds exceeds all others by at least two orders of magnitude. The "Philosophy" cycle which contains "Existence," "Aware-

ness," "Reality," and similar articles accumulates the overwhelming proportion of its references through "Philosophy": 7.37 million of the 7.4 million references are funneled through "Philosophy." Second on the list of

## 3-cycles

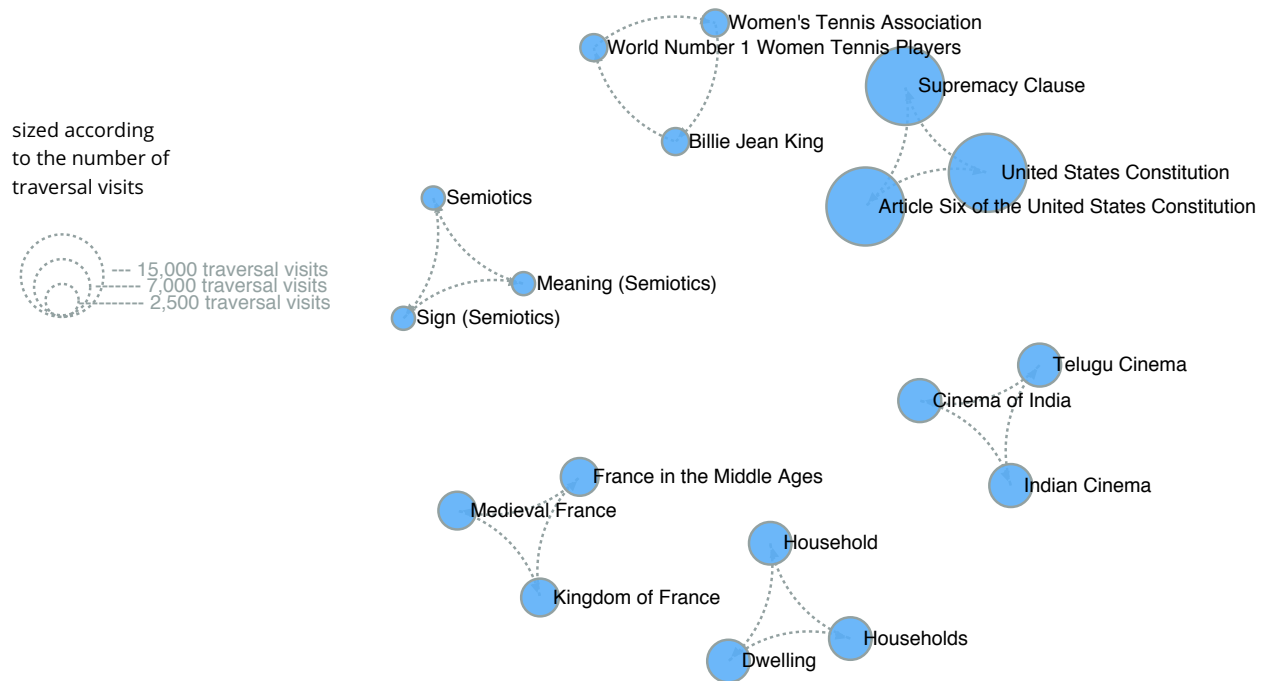


FIG. 11. **highest ranking 3-cycles** We identify groups of three articles whose links form a 3-cycle with the last article in the group linking back to the first article. We rank each 3-cycle by the total number of traversal visits to gauge the most referenced group of three linked articles.

## Articles with the highest traversal funnels

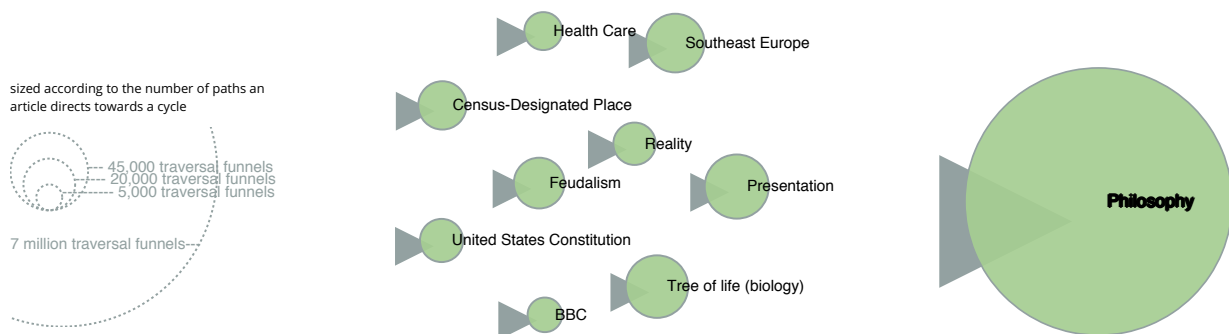


FIG. 12. **Funnels** We represent the highest-ranking articles by the number of traversal funnels to gauge the influence each article exerts in shaping the structure of the FLN. We find "Philosophy" exerts an overwhelming proportion of the influence with other abstract notions and topical concepts ranking next.

highest-ranking articles by traversal funnels is "Presentation" with only 30 thousand paths. Similarly abstract ideas also rank highly such as "Tree of life" (30 thousand), "Reality" (13 thousand), "Jurisdiction" (3 thousand) also appear of the list.

Curiously, many high-ranking articles are remarkably topical, culturally and politically important ideas. For example, "Health Care", a recently high-contested legislative topic appears high on the list—Google trends indicates an uncharacteristic spike in search frequency between August-2009 and February-2010. Other high ranking articles include key historical events such as the "Cold War" or critical scarce resource with recent media discussion such as "Fossil Fuel". The highest-ranking list even includes "Hip Hop," "Cancer," and "Web Page." This coincidence of recent relevance and traversal feed rank suggests the First Link Network measurably represents meaningful relationships not only among ideas, but also to English-speaking society.

As a distribution, we find few articles influence the structure of the FLN. Only 17,821 articles have one or more traversal funnels, leaving an more than 99% with none—most articles are recipients of the references flowing through the articles with at least one traversal funnel. When fit to a log-log linear model we find the 99% of articles with zero traversal funnels forming one regime (with  $\log(\text{rank})$  less than 9). The top regime, corresponding to the 17,821 articles with at least one traversal funnel strongly fits a linear model with an  $r$ -value of  $-0.99$  to yield a power-law exponent of  $-1.08$ . Even within the few articles influencing the structure of the FLN, only a handful of these exert most of the control.

The flow of ideas in the FLN is overwhelmingly influenced by "Philosophy," with a small proportion guided by "Presentation," "Tree of life (biology)," "Southeastern Europe," "Feudalism," "Census-Designated Place," "United States Constitution," "Reality," and "Health Care."

#### G. Article Popularity

Wikipedia released an API to measure article popularity by page views starting November 2015. The page views adds another dimension to our findings by revealing how users view the information on Wikipedia. We measure popularity as the total number of page views in the English edition of Wikipedia in the month of October 2015—the earliest full month for which the data is available. We find the highest ranking 1000 articles by traversal visits have an average of 70 thousand page views in October with high variation: the standard deviation is 1.1 million page views. The number of page views has a skewed distribution with 75% of articles reaching fewer than 73 thousand page views in October. The article for "United States" has the most page views

of the highest ranking 1000 articles by traversal visits with 1.4 million views in October. The next most popular articles are "India," "World War II," and the "United Kingdom" each with 0.7 – 1.03 million page views.

The "Philosophy" article despite outranking every article in traversal visits, has only 240 thousand page views in October compared to the 1.4 million page views for "United States". While "Philosophy" is not itself the most popular article, "Philosophy" serves as an accumulation points for the ideas. Wikipedia users do not login specifically to read about "Philosophy," but the authors relate ideas to "Philosophy" by ultimately anchoring the flow of first links to "Philosophy."

We also analyze article popularity for the highest ranking 1000 articles by traversal funnels. In October, the average page views per article is 22 thousand with a standard deviation of 95 thousand views. The distribution is skewed with 75% of articles reaching fewer than 20 thousand page views. "Halloween" is the most popular article with 2.8 million views in October, likely a result of Halloween falling in the month of October. Other popular articles include "Scientology" (463 thousand views), "Clint Eastwood" (341 thousand views), and the "Cold War" (298 thousand views) although each has significantly fewer views compared to the views for "Halloween." Other standout popular October articles include "24-hour" and "12-hour" clocks likely due to Daylight savings in October and "Marriage" likely due to the drop in the number of weddings in the months following October [21]. "Philosophy" which ranks seventh among the most popular of the highest ranking articles by traversal funnels, appears nowhere in the top 20 by traversal visits. "Philosophy" is a relatively popular article among articles influencing the shape of the network, but less popular among articles with a large accumulation among the flow of first links.

#### IV. REFLECTIONS

The findings here should only be considered within the limitations of their context. We examined only the English version of Wikipedia at a particular moment in time. Furthermore, we only studied the First Link in the main body of each article as a means to related one article to another. Finally, Wikipedia, while the largest collection of human knowledge, is rife with the biases of the many contributing editors [34]. Nevertheless, the findings do reveal generalizable relationships, point to foundational notions, and uncover many curiosities.

Among the curiosities is the multiple appearance of scale-free distributions within the network. The three metrics we developed: path length, traversal visits, and traversal funnels are all marked by scale-free distributions. Few articles have most traversal visits, few paths

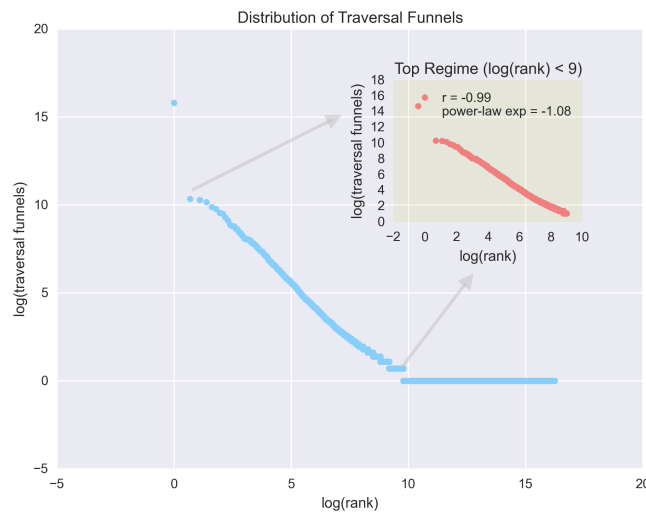


FIG. 13. **Distribution of Traversal Funnels** We fit the distribution to a linear log-log model by considering the log (base 10) transformed rank of each article against log (base 10) transformed the number of traversal funnels. We find two regimes with the top regime ( $\log(\text{rank}) < 9$ ) well-explained by a linear fit, yielding an  $r$ -value of  $-0.99$  and a power-law exponent of  $-1.08$ . The top regime explains the distribution of traversal funnels for the 17,821 articles with at least one traversal funnel. The bottom horizontal regime corresponds to the more than 99% of articles which hold zero traversal funnels.

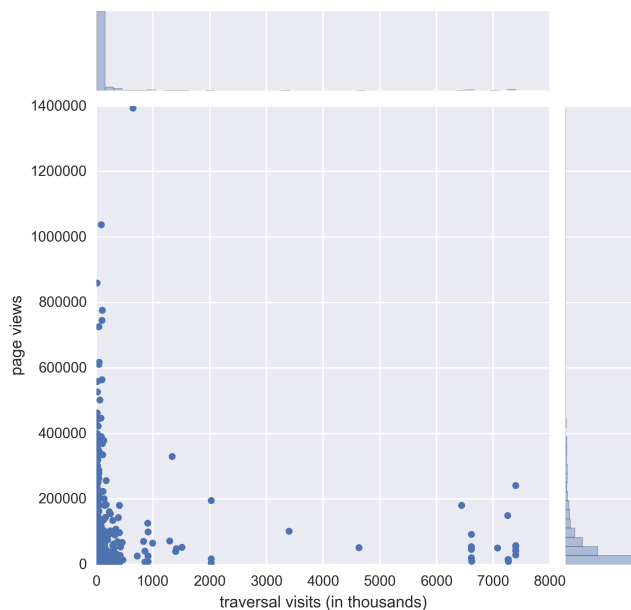


FIG. 14. **Article Popularity (by page views) compared to Traversal Visits for Highest ranking 1000 articles** We use the total number of page views provided by Wikipedia for the month of October 2015 to compare each article's popularity to traversal visits. While the most popular articles do not necessarily correspond to the articles with the highest number of traversal visits, the variation in popularity appears to decrease as the number of traversal visits increases.

have an exceptionally long path length, and even fewer articles are responsible for funneling most paths. When measured against the traversal funnels, "Philosophy" emerges as an exceptional article by orders of magnitude. Nevertheless, many other foundational ideas emerged naturally within the First Link Network. Basins around "Community", "State", and "Science" reveal a founda-

tional structure within the network. More curious is the emergence of recently prominent political and economics topics such as "Fossil Fuel" and "Health Care" within the highest ranking funnels. Wikipedia seems to reflect not only timeless foundations, but also the topical (at least within English speaking society).

Future work could examine other language versions

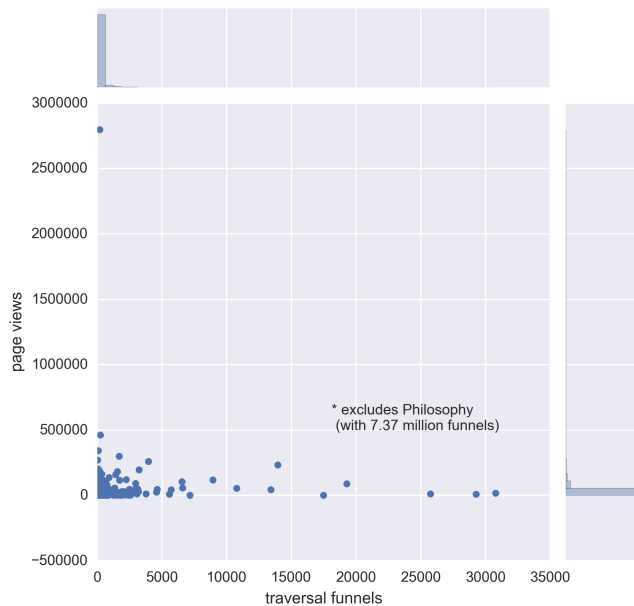


FIG. 15. **Article Popularity (by page views) compared to Traversal Funnels for Highest ranking 1000 articles**

We use the total number of page views provided by Wikipedia for the month of October 2015 to compare each article's popularity to traversal funnels. In the top left is "Halloween" with a significantly greater number of page views in the month of October.

of Wikipedia for potentially telling cultural or regional differences as well as expand the network to more than the First Link. These findings also form the basis for the creation of a taxonomy where every idea, event, or object sits within a hierarchy of connected notions. The taxonomy would extend a traditional word thesaurus beyond mere synonyms to a related hierarchy of concepts. Applications could range from an enhanced thesaurus of ideas to psychological insights into how humans form associations. Specifically, an ever-evolving reference of related hierarchical concepts can be applied to search engine algorithms or natural language processing.

#### ACKNOWLEDGMENTS

Thank you to my friend RJ for pointing out the Reddit post [33] describing how the majority of links lead to "Philosophy," inspiring this research.

## V. APPENDIX

### A. Constructing The First Link Network

To map Wikipedia's First Link Network, we use the freely-available XML dump of the English edition of Wikipedia. Rather than rely on a sample of articles from which to generalize, we opted to process the entirety of Wikipedia, eliminating any statistical error due to sampling. We analyze the snapshot provided on November 2014, representing the state of Wikipedia at the time. The November raw dump consists of 11 million articles: 4.7 million unique articles along with redirects and disambiguations. Knowing Wikipedia is an ever-evolving project with 10 edits every second and 750 new articles per day on average [35], our aim is to characterize the dynamics of the First Link Network, not record a particular link between one article and another.

Wikipedia renders and stores articles in MediaWiki markup, a markup language with syntax and keywords to format and mark elements in a page. Along with special syntax for links, MediaWiki markup includes templates for audio files, images, and side-bar information. While a human could accurately identify the First Link, to map the entire First Link Network of 11 million articles, we needed to programmatically untangle the body text from side-bar, header box, and invalid link elements.

While some libraries exist for MediaWiki Markup, 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. Consequently, we developed an algorithm for parsing the First Link in the XML version of each article. Our parsing algorithm aimed to: 1) accurately identify the First Link among other page elements, and 2) efficiently do so—that is without needing for several passes through the data. To process an article in a single pass, we developed a hierarchical system of flags:

The algorithm loops in three-character chunks to account for potentially nested elements, shifting by one character steps through the article markup. If any markup triggers for a flag are detected, a flag is raised.

Once a flag is raised, we stop processing and proceed to the next character until the flag's closing markup. A First Link is identified only if Flags 1, 2, and 3 are all off. In this case, the entire link is retrieved. We then confirm the link is valid by filtering for MediaWiki keywords indicating external page or other projects as well as common file extensions for images, audio files, and the like [36]. The First Link of an article is then the earliest valid link with unraised flags.

To process the entirety of Wikipedia, we distributed the parsing and processing of the XML dump across 112 cores of the UVM supercomputer cluster [19]. We then joined the results to form a hash table containing every Wikipedia article and its corresponding First Link. The resulting network map is the basis of our analysis.

- 
- [1] Iba, Takashi, et al. "Analyzing the creative editing behavior of Wikipedia editors: Through dynamic social network analysis." *Procedia-Social and Behavioral Sciences* 2.4 (2010): 6441-6456.
  - [2] Holman Rector, Lucy. "Comparison of Wikipedia and other encyclopedias for accuracy, breadth, and depth in historical articles." *Reference services review* 36.1 (2008): 7-22.
  - [3] Giles, Jim. "Internet encyclopaedias go head to head." *Nature* 438.7070 (2005): 900-901.
  - [4] Halavais, Alexander, and Derek Lackaff. "An analysis of topical coverage of Wikipedia." *Journal of Computer-Mediated Communication* 13.2 (2008): 429-440.
  - [5] Hill, Benjamin Mako, and Aaron Shaw. "The Wikipedia gender gap revisited: characterizing survey response bias with propensity score estimation." *PloS one* 8.6 (2013): e65782.
  - [6] Williams, Jake Ryland, Clark, Eric M., Bagrow, James P., Danforth, Christopher M. & Dodds, Peter Sheridan (2015). Identifying missing dictionary entries with frequency-conserving context models. *Phys. Rev. E*, 92, 042808.
  - [7] Banerjee, Somnath, Krishnan Ramanathan, and Ajay Gupta. "Clustering short texts using wikipedia." *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2007.
  - [8] Gabrilovich, Evgeniy, and Shaul Markovitch. "Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis." *IJCAI*. Vol. 7. 2007.
  - [9] Cucerzan, Silviu. "Large-Scale Named Entity Disambiguation Based on Wikipedia Data." *EMNLP-CoNLL*. Vol. 7. 2007.
  - [10] Clauson, Kevin A., et al. "Scope, completeness, and accuracy of drug information in Wikipedia." *Annals of Pharmacotherapy* 42.12 (2008): 1814-1821.
  - [11] Wikipedia contributors. "Wikipedia:Statistics." Wikipedia, The Free Encyclopedia. Wikipedia, The Free Encyclopedia, 9 Nov. 2015. Web. 13 Nov. 2015.
  - [12] Page, Lawrence, et al. "The PageRank citation ranking: bringing order to the Web." (1999).
  - [13] Chakrabarti, Soumen, Mukul Joshi, and Vivek Tawde. "Enhanced topic distillation using text, markup tags, and hyperlinks." *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2001.
  - [14] Kamps, Jaap, and Marijn Koolen. "Is Wikipedia link structure different?." *Proceedings of the second ACM international conference on Web search and data mining*. ACM, 2009.
  - [15] Bolton, Martha Brandt. "The Taxonomy of Ideas in Locke's *Essay*", *The Cambridge Companion to Locke's 'Essay Concerning Human Understanding'*. Ed. Lex Newman.. 1st ed. Cambridge: Cambridge University Press, 2007. 67-100. Cambridge Companions Online. Web. 13 November 2015. <http://dx.doi.org/10.1017/CCOL0521834333.004>.
  - [16] Studtmann, Paul, "Aristotle's Categories", *The Stanford Encyclopedia of Philosophy* (Summer 2014 Edition), Edward N. Zalta (ed.), <http://plato.stanford.edu/archives/sum2014/entries/aristotle-categories/>
  - [17] Smith, Kurt, "Descartes' Theory of Ideas", *The Stanford Encyclopedia of Philosophy* (Spring 2014 Edition), Edward N. Zalta (ed.), <http://plato.stanford.edu/archives/spr2014/entries/descartes-ideas/>
  - [18] What is the Historical Thesaurus of the OED - Oxford English Dictionary (Oxford English Dictionary) <http://public.oed.com/historical-thesaurus-of-the-oed/what-is-the-historical-thesaurus-of-the-oed/>
  - [19] The authors acknowledge the Vermont Advanced Computing Core which is supported by NASA (NNX 06AC88G), at the University of Vermont for providing High Performance Computing resources that have contributed to the research results reported within this paper. <http://www.uvm.edu/vacc>
  - [20] Page Views for Wikipedia, Non-mobile site, Normalized (Page Views for Wikipedia, Non-mobile site, Normalized) <https://stats.wikimedia.org/EN/TablesPageViewsMonthly.htm>
  - [21] Wedding statistics in the United States (Wedding statistics in the United States) <http://www.soundvision.com/>

## Parsing Algorithm





-  1: inside Wikimedia template?  
trigger: {{ }}
-  2: inside <ref>, <div>?
-  3: inside ()?
-  valid link to Wikipedia article? ☒

FIG. 16. **Parsing Algorithm of Wikipedia's XML dump** The highest flag in the hierarchy indicates a Wikimedia template used to mark an element in the side bar, display an image, link to an external file, or another Wikimedia project outside of Wikipedia. Next, to catch any remaining elements outside the main body we have a second flag for <ref>, <div> elements. Finally, we identify parenthesis to ensure we do not capture a link to a pronunciation key.

- article/wedding-statistics-in-the-united-states
- [22] Elmacioglu, Ergin, and Dongwon Lee. "On six degrees of separation in DBLP-DB and more." *ACM SIGMOD Record* 34.2 (2005): 33-40.
  - [23] Princeton University "About WordNet." WordNet. Princeton University. 2010. <http://wordnet.princeton.edu>
  - [24] Newman, Mark EJ. "The structure and function of complex networks." *SIAM review* 45.2 (2003): 167-256.
  - [25] Eubank, Stephen, et al. "Modelling disease outbreaks in realistic urban social networks." *Nature* 429.6988 (2004): 180-184.
  - [26] Newman, Mark EJ, Duncan J. Watts, and Steven H. Strogatz. "Random graph models of social networks." *Proceedings of the National Academy of Sciences* 99.suppl 1 (2002): 2566-2572.
  - [27] Horton, Robert E. "Erosional development of streams and their drainage basins; hydrophysical approach to quantitative morphology." *Geological society of America bulletin* 56.3 (1945): 275-370. APA
  - [28] Dodds, Peter Sheridan, and Daniel H. Rothman. "Unified view of scaling laws for river networks." *Physical Review E* 59.5 (1999): 4865.
  - [29] Garlaschelli, Diego, Guido Caldarelli, and Luciano Pietronero. "Universal scaling relations in food webs." *Nature* 423.6936 (2003): 165-168.
  - [30] brain of mat kelcey (brain of mat kelcey). "Do all first links lead to Philosophy?" <http://matpalm.com/blog/2011/08/13/wikipedia-philosophy/>
  - [31] User:Ilmari Karonen/First link (Wikipedia) [https://en.wikipedia.org/wiki/User:Ilmari\\_Karonen/First\\_link](https://en.wikipedia.org/wiki/User:Ilmari_Karonen/First_link)
  - [32] xkcd: Extended Mind (xkcd: Extended Mind) <http://xkcd.com/903/>
  - [33] Almost every Wikipedia page links to Philosophy if you keep clicking the first link in every article. [https://www.reddit.com/r/InternetIsBeautiful/comments/35asgt/almost\\_every\\_wikipedia\\_page\\_links\\_to\\_philosophy/](https://www.reddit.com/r/InternetIsBeautiful/comments/35asgt/almost_every_wikipedia_page_links_to_philosophy/)
  - [34] Wagner, Claudia, et al. "It's a Man's Wikipedia? Assessing Gender Inequality in an Online Encyclopedia." *arXiv preprint arXiv:1501.06307* (2015).
  - [35] Wikipedia:Statistics <https://en.wikipedia.org/wiki/Wikipedia:Statistics>
  - [36] MediaWiki: "Help: Templates" <https://www.mediawiki.org/wiki/Help:Templates>