Connecting every bit of knowledge:
The Structure of Wikipedia's first link network

Mark Ibrahim,[1, *] Christopher M. Danforth,[1, †] and Peter Sheridan Dodds[1, ‡]

[1]*Department of Mathematics & Statistics, Computational Story Lab,*
Vermont Complex Systems Center, Vermont Advanced Computing Core, The University of Vermont, Burlington, VT 05401.

Apples, porcupines, and the most obscure Dylan song—is every topic 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, Wikipedia is a web of naturally emerging relationships. By following the first link in an article, we connect entries to form a directed network: Wikipedia's First Link Network. Here, we study the English edition of Wikipedia's First Link Network for insight into how the many topics, ideas, people, objects, and events are related and organized. We algorithmically parse all 4.7 million articles to construct a map of Wikipedia's First Link Network. We traverse every possible path through the network, measuring the accumulation of first links, path lengths, basins, cycles, and the influence a particular article exerts in shaping the network. We discover many scale-free distributions, find Philosophy at a salient center, and uncover a flow from specific to general culminating 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 Wikipedia's ever growing store of knowledge.

TODO:

- fix end of second paragraph in Results Section
  * power-law biz and total peter is asking for

- fix exponent reporting (with alpha)
  * figure out what gamma size freq means

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 [10, 11]. All this information does not lie dormant. Wikipedia is the sixth most visited site in the world, surpassing 18 billion page views and 26 million edits in a single month [21, 36].

Wikipedia has naturally become 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 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. A hyperlink from one Wikipedia article to another indicates a relationship between the two articles [14]. Naturally, the notion that hyperlinks convey information about the content of a page has proved enormously successful in multiple domains from search engine algorithms such as PageRank [12] to topic classification [13]. We treat a hyperlink as a mechanism connecting two topics.

The authors of a Wikipedia article choose where and whether to include a reference to another Wikipedia article in the HTML markup. For example, the authors of the "Train" article chose "Rail Transport," "Cargo," and "Passengers" as pertinent articles to reference in describing trains [15] . Here we focus on the first hyperlink to another Wikipedia article, which marks the earliest moment in the body text where the authors chose to connect the information discussed to that of another article. By following the first hyperlink to another article in the English edition of Wikipedia as of November, 2014, 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 [31–33] —we holistically study Wikipedia's FLN as a map relating areas of human knowledge. The map

---

[*] mark.s.ibrahim@uvm.edu
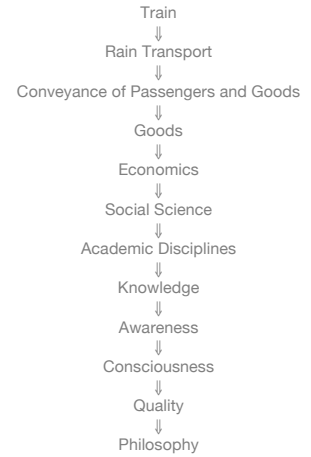[†] chris.danforth@uvm.edu
[‡] peter.dodds@uvm.edu
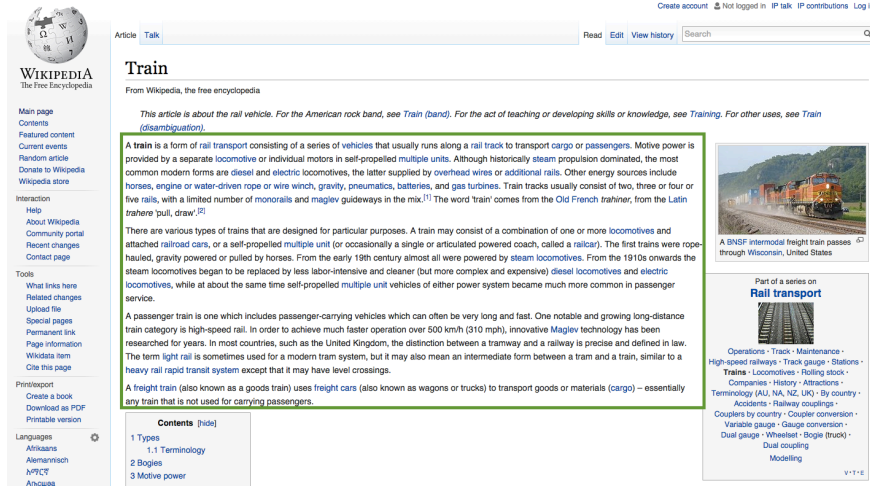
FIG. 1. **First Link Path For "Train."** We follow the First Link to another Wikipedia article in the main body of the article—the area inside the green rectangle, which excludes side bar elements, the navigation bar and title. In this example, the First Link to another Wikipedia article is "Rail Transport." We can again select the First Link on the "Rail Transport" article, repeating the process to form a path of First Links.

is a hierarchical taxonomy where, for example, "Train" links to a parent node, "Rail Transport," while many child nodes, "Steel" and "Horsepower," link to "Train." Unlike previous taxonomies created by individuals [16–18] or a select group of individuals [19], the organization of ideas in the FLN does not emerge from a centralized effort. Rather each link in the FLN emerges from the choice of each author and ensuing edits. Collectively, the FLN is a wealth of relations among inventions, places, figures, objects, and events across space and time.

Our goal is to study the structure of the FLN for insight into how the information on Wikipedia is organized and related. 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, in-degree, 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

An essential feature of a directed network's structure is the degree distribution [25]. The degree distribution has been used to study many phenomena from disease outbreak [26] to the dynamics of social networks [27]. The in-degree distribution in the FLN describes how many first links point to a particular article. We measure the in-degree in the FLN to gauge the number of articles referencing a particular article. Articles with zero in-degree have no references—they are outer leaves in the FLN. Similar to PageRank, the in-degree provides a way to rank the articles in Wikipedia, allowing us to see whether authors tend to reference all articles equally or only a few.

Our greater interest is lies in the richer dynamics of how many articles fit together—how links create a flow through the FLN. We construct a path through the FLN with each article leading to the next through 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. As shown in Fig. 1, the path starting at "Train" contains "Goods," "Economics," "Social Science," ultimately leading to "Philosophy".

Starting at each article, we construct a path through the FLN, ultimately mapping a flow of connections among articles. 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.

We use the notion of a path through the FLN to describe the structure of the FLN as a flow relating topics. Grounded in the dynamics of river networks [28], we characterize features of the FLN as a flow through first links. Previous studies have used flow to characterize the structure of river networks [29] and describe
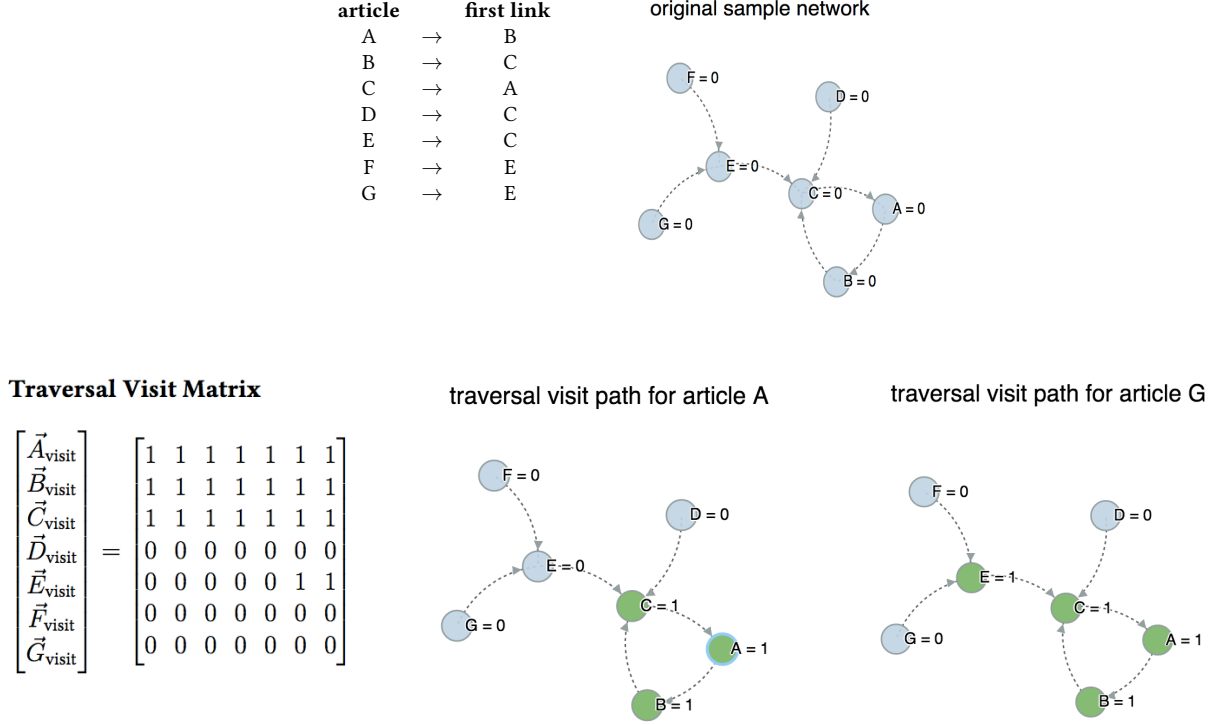
Traversal Algorithm

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 is 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: $\sum_{i=1}^{7} A_{\text{visit, i}} = 7$.

the organization of food systems as transportation networks [30]. With paths to mark flow through the FLN, we develop 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.

The first metric we develop quantifies the accumulation of first links. 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 continue until the first link is retraced or is invalid—defined here to mean outside of Wikipedia. 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 for 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—equivalent to basin area in geomorphology.

We can characterize the paths in the FLN as a matrix with each column corresponding to a path. In our sample network (Fig. 2), the path starting at article A is the first column in the traversal visits matrix. An entry of
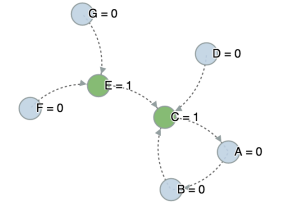

FIG. 3. **Traversal Funnel Algorithm on a sample network.** The algorithm for traversal funnels is identical to the previous algorithm for traversal visits with one alteration: the path ends at the start of a cycle to distinguish articles directing a path into a cycle from articles that simply happen 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 $\sum_{i=1}^{7} E_{\text{funnel, i}} = 2$.

1 indicates the path contains a given article and 0 indicates the path does not. To compute the number of

traversal visits for an article, we sum the corresponding row in our matrix. The traversal visits matrix for our Wikipedia dataset consists of 121 million entries to account for each path through the FLN.

By measuring the number of first links between two articles, we obtain an additional piece of information we call *path length.* We can compute the length of a path by summing along a column in our traversal visits matrix. In our Wikipedia dataset, the sum of all path lengths is 232 million first links. The path length describes how closely related topics are. The greater the number of first links separating two articles, the greater the number of ideas needed to relate their corresponding ideas. Although "Train" is related to "Economics" for example, there are several 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."

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 (Fig. 2) a 3-cycle exists among nodes A, B, and C. We can readily identify the types of cycle structures within the FLN and rank each 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 and identifies the paths to a particular article.

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 on 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 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 into a particular cycle. 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 direct influence themselves.

By studying the FLN not only as collection of directly linked pairs of articles, but as a flow, we build a powerful arsenal of information. From accumulated ref-
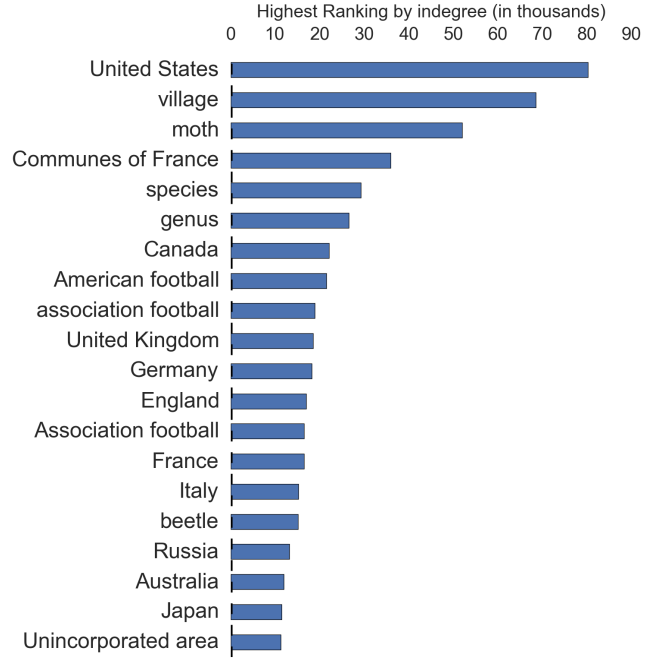

FIG. 4. **Highest Ranking Articles by in-degree.**
We rank each article by the number of direct first links to the article (in-degree). The highest-ranking articles tend to represent abstraction for concepts comprised of various facets.

erences, cycles, and basins, we can measure how the many articles in Wikipedia are organized and related.

III. RESULTS

A. Degree Distribution

We rank all 11 million articles by in-degree to find the "United States" with 80,249 direct first links as the most referenced Wikipedia article (Fig. 4). 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 with a common culture, language, or geographical proximity; sports teams describe an ever changing collection of sports players often associated with a cultural identity or a geographical region. Since abstractions such as nations and teams are inherently comprised of many parts, authors reference the abstraction when describing a part. In describing the New England Patriots or the New York Giants for example, "American Football" is the natural abstraction, which anchors many specific teams.
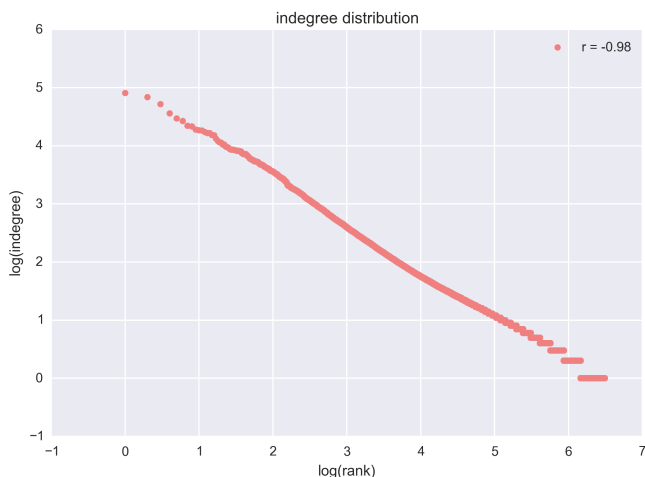
FIG. 5. **FLN Degree Distribution.**
We construct a log-log plot and fit our results with a linear model. The result is a an excellent fit with r = -0.98, yielding a power law exponent of -0.79. The distribution appears to be scale free.

"Philosophy" and other philosophical concepts are not among the highest-ranking articles by in-degree. "Philosophy" has an in-degree of only 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 below), the accumulation is not the result of many articles directly referencing "Philosophy." Instead, flows towards Philosophy as the ultimate anchor when generalizing from specific articles to broader notions.

The FLN's in-degree exhibits a decaying power-law distribution where a few articles receive most direct first references, while most articles receive few or none. The average in-degree 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 with a linear model on a log-log scale (log(rank) versus log(in-degree)), the model's r value is -0.98 suggesting a strong log-log linear fit with a power law exponent of -0.79.

B. Depth of the FLN

We first seek to describe the depth of the FLN: how many links does a connection of among articles span? We find the longest path length is 365, corresponding to the yearly calendar of Orthodox Liturgics. Each day's
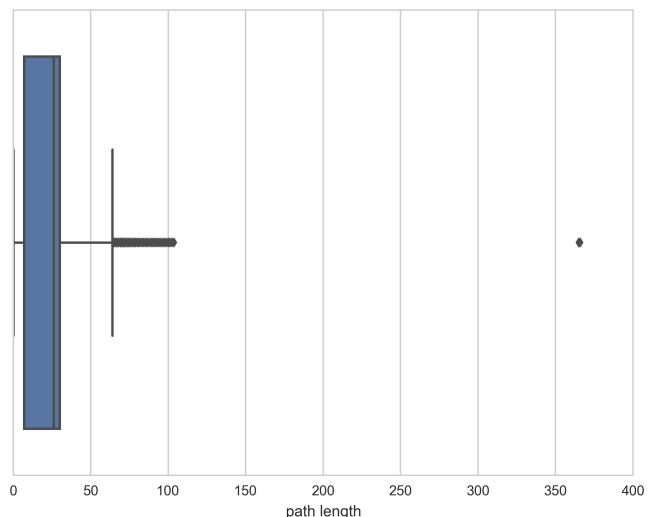

FIG. 6. **Path Length Distribution.**
The median network depth measured by the number of first links in a path is 29. The 365 cycle for "Orthodox" liturgics is the outlier to the right, while other historical articles about "Scotland," "UK" and so on are slightly outside the third quartile. More than 75% of articles have path lengths between 0 and 50 links.

Liturgics links to the next day's. On the last calendar day, the last article simply links back to January 1, forming a 365-cycle (see discussion of cycles). We also find 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. In general, the longest paths connect temporally organized ideas.

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 of 4—path length is typically between 26 and 29 articles. 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).

C. Traversal Visits

As a distribution, the number of traversal visits by article appears to follow a decaying power law. The majority of articles have fewer than 30 traversal visits and first link references accumulate at a few articles. Specifically, 99.76% of articles have fewer than 100 traversal visits; nearly 80% have none. Meanwhile, the
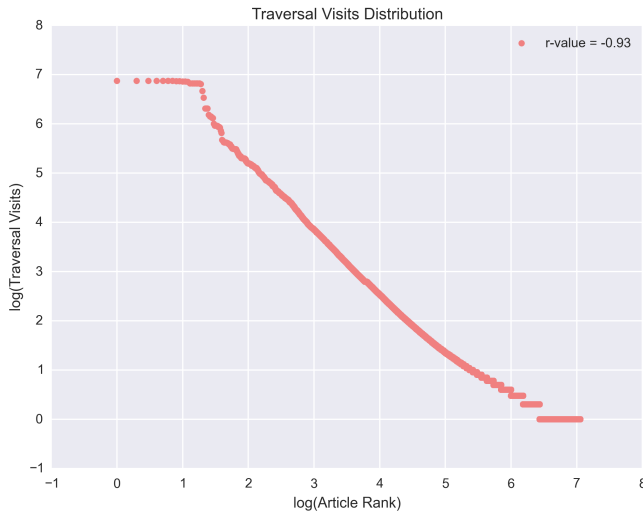
FIG. 7. **Distribution of Traversal Visits.**
We fit the distribution to a linear log-log model by considering the log (base 10) transformed rank of each article against log (base 10) transformed the number of traversal visits. The model explains 86% of the variation in the data, yielding an r-value of -0.93 and a power law exponent of -0.64. The horizontal flattening around the highest ranking articles is a result of the cyclic structure (see discussion on cycles).

highest ranking 30 articles have an extremely disproportionate number of traversal visits.

In Fig., we show a log-log plot for the entire dataset:traversal visits against rank. We observe a strong linear fit, with a Pearson's Correlation Coefficient of -.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.

D. Network Cycles

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 Fig. 9).

Outside of the highest ranking 2-cycles, the typical 2-cycle signals a connection between distinct, yet very closely related ideas. We also observe link patterns such as inventor to product ("Voere" to "VEC-91"), event to organizer ("Poetry Bus Tour" to "Weave Books"),
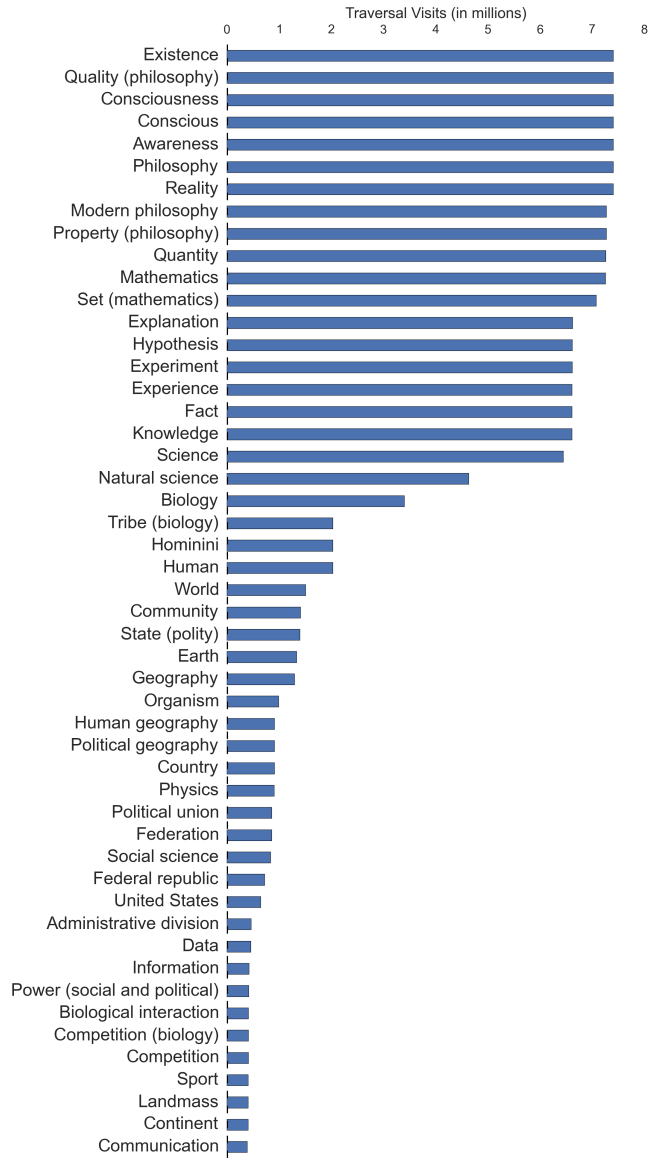

FIG. 8. **Highest ranking articles by number of traversal visits.**
We compute the number of traversal visits for each article in the FLN (see Traversal Algorithm section for details). In doing so, we can rank each article by the accumulation of first links. Articles with a greater number of traversal visits mark greater points of first link accumulation. The highest ranking articles by traversal visits reveal where the greatest accumulation occurs.

and book to author ("Anatomy of Britain" to "Anthony Sampson").

Similarly, 3-cycles capture 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 10). Once we extend our cycle size beyond a length of 6 however, "Philosophy" along with the remaining
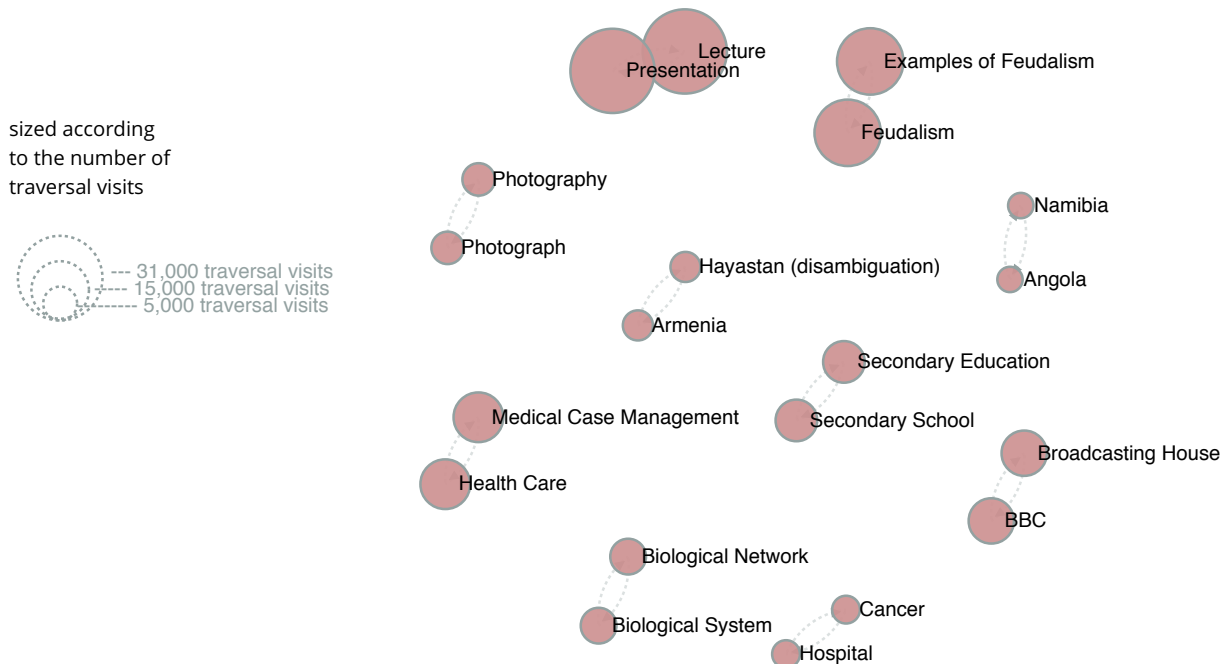
2-cycles

FIG. 9. **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. The area of the circles in the figure indicates the number of traversal visits. We find 2-cycles often capture synonyms or articles representing nearly the same concepts as opposed to distinct ideas.

list of high ranking articles by traversal visits dominate. The longest cycle in the network spans 365 articles of Eastern Orthodox Liturgics for each calendar day. 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 for each article in 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." 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 concepts 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
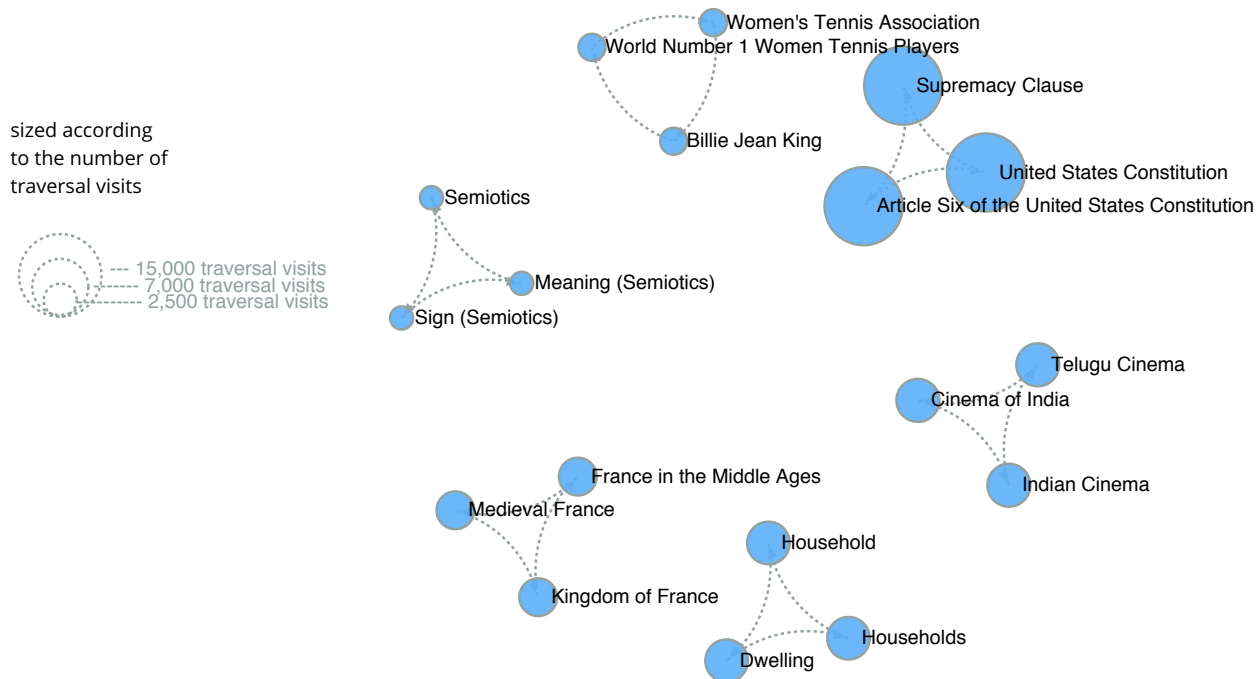
3-cycles

FIG. 10. **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. The area of the circles in the figure indicates the number of traversal visits.

with "Democracy," another beginning with "Constitution," another at "Dictatorship," and so on. The ideas build from specific means of organizing a community (or society) and then build 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".

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" to be by far the highest-ranking

article with 7.37 million paths (see figure 11). Of any article, the number of traversal funnels Philosophy holds exceeds all others by more than two orders of magnitude. The "Philosophy" cycle which contains "Existence," "Awareness," "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 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), and "Jurisdiction" (3 thousand).

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

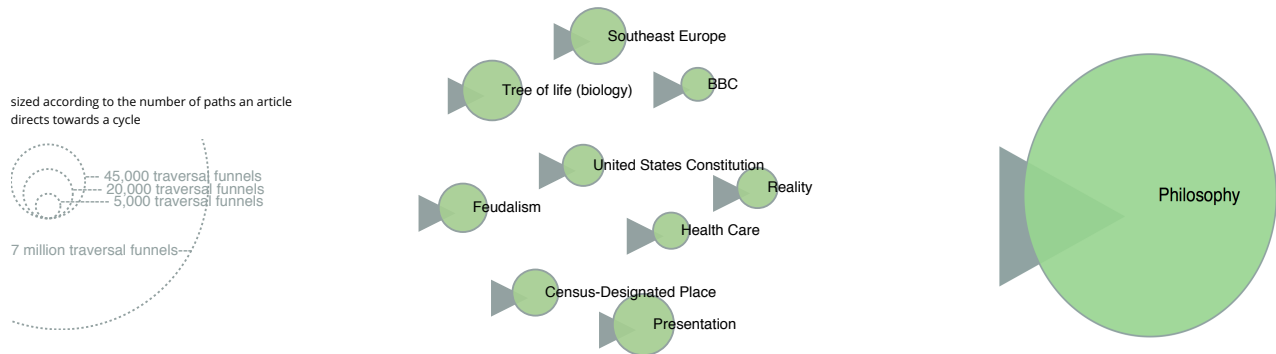## Articles with the highest traversal funnels

FIG. 11.   **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. The area of each circle is sized in proportion to the number of traversal funnels each article has. We find "Philosophy" exerts an overwhelming proportion of the influence, with other abstract notions and topical concepts ranking next.
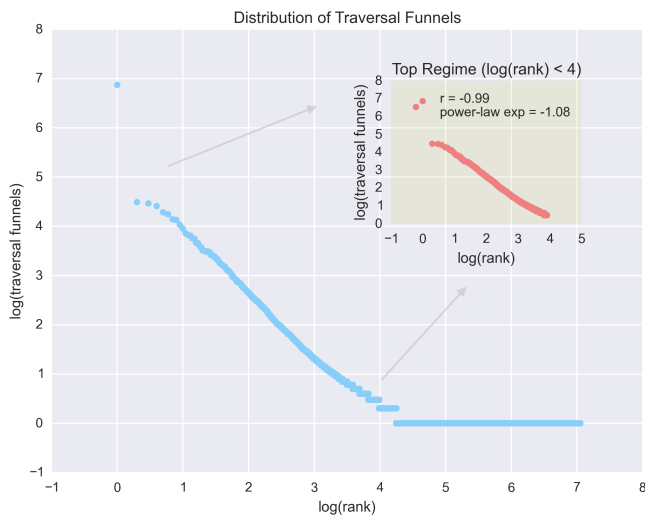

FIG. 12.   **Distribution of Traversal Funnels.** We fit the distribution with a linear log-log model by considering the log (base 10) transformed rank of each article against log (base 10) transformed the number of traversal funnels. We find two regimes with the top regime (log(rank) < 4) well-explained 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.

also includes "Hip Hop," "Cancer," and "Web Page".

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 form one regime (with log(rank) less than 4). 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. The resulting power-law exponent is −1.08. Even within the few articles influencing the structure of the FLN, only a handful of these exert most of the control.

G.   Article Popularity

Wikipedia released an API to measure article popularity by page views starting November 2015. Page views add 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 popu-
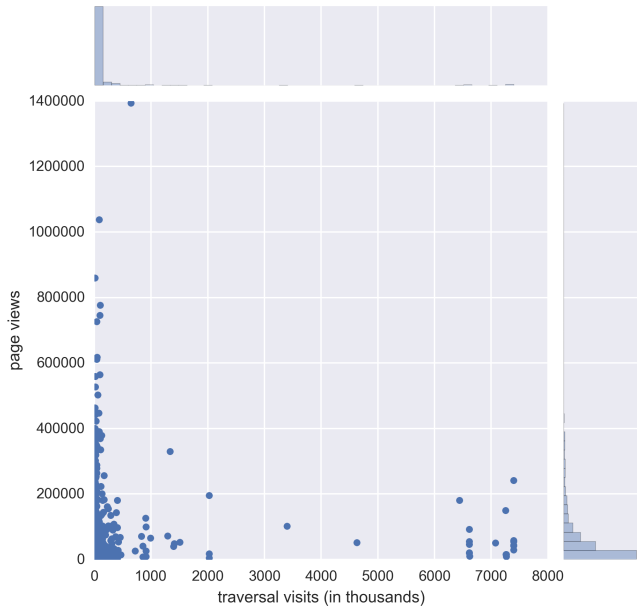
FIG. 13. **Article Popularity (by page views) compared to Traversal Visits for Highest ranking 1000 articles.** We use the total number of page views provided by Wikipedia for the month of October 2015 to compare each article's popularity to traversal visits. While the most popular articles do not necessarily 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.
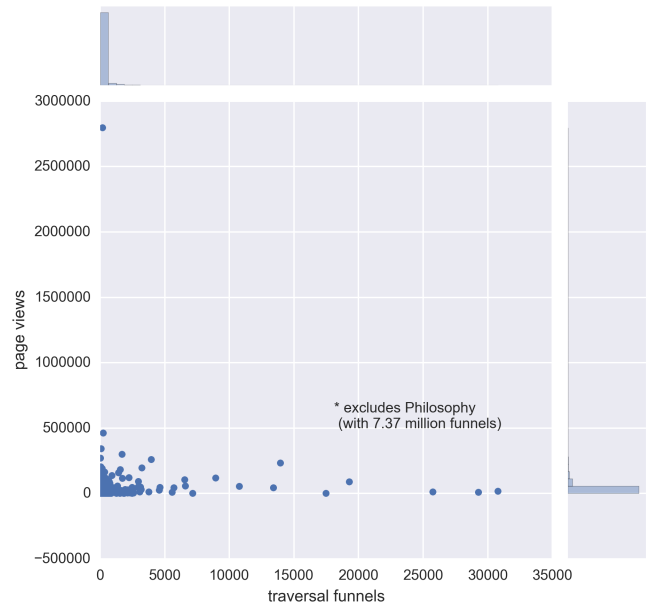

FIG. 14. **Article Popularity (by page views) compared to Traversal Funnels for Highest ranking 1000 articles.**
We use the total number of page views provided by Wikipedia for the month of October 2015 to compare each article's popularity to traversal funnels. In the top left is "Halloween" with a significantly greater number of page views in the month of October.

lar 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 by traversal visits, has only 240 thousand page views in October compared to the 1.4 million page views for "United States".

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 [22]. "Philosophy" which ranks seventh among the most popular of the highest ranking articles by traversal funnels, appears no where 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 highly-ranked articles by accumulation.

IV. CONCLUDING REMARKS

The findings here should only be considered within the limitations of their context. We examined only the English editiion of Wikipedia at a particular moment in time, considering only the first link in in 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 [35] . Nevertheless, the findings do reveal generalizable relationships and point to foundational notions.

Among our observations is the multiple appearance of scale-free distributions within the network. Few articles have most traversal visits, few paths 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 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 network of ideas to psychological insights into how humans form associations. Specifically, an ever-evolving reference of related hierarchical concepts could be applied to search engine algorithms or natural language processing.

ACKNOWLEDGMENTS

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 [36], 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 untangled 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 [37]. 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

Parsing Algorithm

🚩 1: inside Wikimedia template?
   trigger: {{ }}

   🚩 2: inside <ref>, <div>?

      🏴 3: inside ()?
         ➡ valid link to Wikipedia article? ☑

FIG. 15. **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 Wikipeida. 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.

cores of the UVM supercomputer cluster [20] 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 ComputerMediated 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] Wikipedia contributors. "Train." Wikipedia, The Free Encyclopedia. Wikipedia, The Free Encyclopedia, https://en.wikipedia.org/wiki/Train

[16] Bolton, Martha Brandt. "The Taxonomy of Ideas in Lockes *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.

[17] 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/

[18] 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/

[19] 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/

[20] 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`

[21] Page Views for Wikipedia, Non-mobile site, Normalized (Page Views for Wikipedia, Non-mobile site, Normalized) `https://stats.wikimedia.org/EN/TablesPageViewsMonthly.htm`

[22] Wedding statistics in the United States (Wedding statistics in the United States) `http://www.soundvision.com/article/wedding-statistics-in-the-united-states`

[23] Elmacioglu, Ergin, and Dongwon Lee. "On six degrees of separation in DBLP-DB and more." ACM SIGMOD Record 34.2 (2005): 33-40.

[24] Princeton University "About WordNet." WordNet. Princeton University. 2010. `http://wordnet.princeton.edu`

[25] Newman, Mark EJ. "The structure and function of complex networks." SIAM review 45.2 (2003): 167-256.

[26] Eubank, Stephen, et al. "Modelling disease outbreaks in realistic urban social networks." Nature 429.6988 (2004): 180-184.

[27] 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.

[28] 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

[29] Dodds, Peter Sheridan, and Daniel H. Rothman. "Unified view of scaling laws for river networks." Physical Review E 59.5 (1999): 4865.

[30] Garlaschelli, Diego, Guido Caldarelli, and Luciano Pietronero. "Universal scaling relations in food webs." Nature 423.6936 (2003): 165-168.

[31] brain of mat kelcey (brain of mat kelcey). "Do all first links lead to Philosophy?" `http://matpalm.com/blog/2011/08/13/wikipedia-philosophy/`

[32] User:Ilmari Karonen/First link (Wikipedia) `https://en.wikipedia.org/wiki/User:Ilmari_Karonen/First_link`

[33] xkcd: Extended Mind (xkcd: Extended Mind) `http://xkcd.com/903/`

[34] 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/`

[35] Wagner, Claudia, et al. "It's a Man's Wikipedia? Assessing Gender Inequality in an Online Encyclopedia." arXiv preprint arXiv:1501.06307 (2015).

[36] Wikipedia:Statistics `https://en.wikipedia.org/wiki/Wikipedia:Statistics`

[37] MediaWiki: "Help: Templates" `https://www.mediawiki.org/wiki/Help:Templates`