Journal of Computational Science 19 (2017) 21–30

Contents lists available at ScienceDirect

Journal of Computational Science

journal homepage: www.elsevier.com/locate/jocs

Connecting every bit of knowledge: The structure of Wikipedia's First Link Network

Mark Ibrahim*, Christopher M. Danforth, Peter Sheridan Dodds

Department of Mathematics & Statistics, Computational Story Lab, Vermont Complex Systems Center, Vermont Advanced Computing Core, The University of Vermont, Burlington, VT 05401, United States

ARTICLE INFO

Article history:
Received 2 May 2016
Received in revised form
21 November 2016
Accepted 5 December 2016
Available online 19 December 2016

Keywords:
Information networks
Graph algorithms
Complex systems

ABSTRACT

Apples, porcupines, and the most obscure Bob 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 each article, we algorithmically construct a directed network of all 4.7 million articles: Wikipedia's First Link Network. Here, we study the English edition of Wikipedia's First Link Network for insight into how the many articles on inventions, places, people, objects, and events are related and organized.

By traversing every path, we measure the accumulation of first links, path lengths, groups of path-connected articles, and cycles. We also develop a new method, traversal funnels, to measure the influence each article exerts in shaping the network. Traversal funnels provide a new measure of influence for directed networks without spill-over into cycles, in contrast to traditional network centrality measures. Within Wikipedia's First Link Network, we find scale-free distributions describe path length, accumulation, and influence. Far from dispersed, first links disproportionately accumulate at a few articles—flowing from specific to general and culminating around fundamental notions such as Community, State, and Science. Philosophy directs more paths than any other article by two orders of magnitude. We also observe a gravitation toward topical articles such as Health Care and Fossil Fuel. These findings enrich our view of the connections and structure of Wikipedia's ever growing store of knowledge.

1. Introduction

Wikipedia is a towering achievement of the Internet age. 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 [1,2]. Demonstrating its relevance to modern inquiry, Wikipedia is the sixth most visited site in the world, surpassing 18 billion page views and 10 million edits in January 2013 alone [3,4].

Wikipedia has naturally become the object of many studies. Researchers have examined the cultural dynamics among editors [5,6], the accuracy of the content relative to traditional encyclopedias [7,8], the topics covered [9], and bias against portions of the population [10]. Wikipedia's content has also proven to be a powerful tool. Researchers have used Wikipedia to identify missing dictionary entries [11], cluster short text [12], compute semantic relatedness [13], and disambiguate meaning [14].

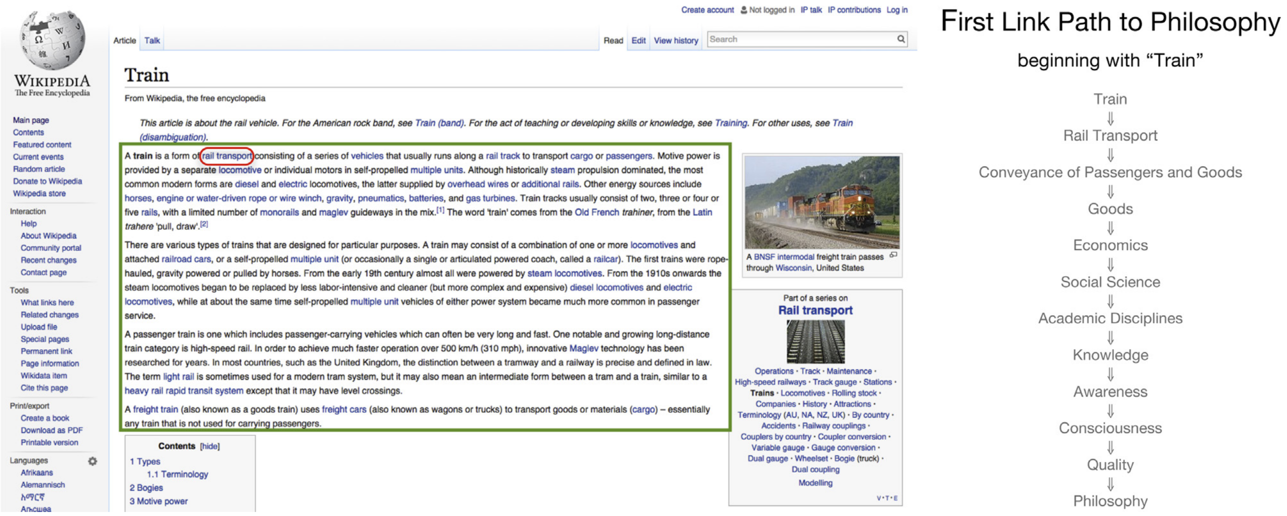
While these many studies have dissected and fruitfully applied Wikipedia's content, here we examine the connections among the many articles. A hyperlink from one Wikipedia article to another naturally indicates a relationship between the two articles [15]. The notion that hyperlinks convey information about the content of a page has proved enormously successful in multiple domains from search engine algorithms such as PageRank [16] to topic classification [17]. Here, 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, as of November 2014, the authors of the "Train" article had collectively chosen "Amtrak's Acela Express," "steam," and "head-end power" among others as relevant articles to reference in describing "Train" [18].

By focusing our attention on the main body of an article—excluding elements in side bars and headings—we attempt to

* Corresponding author.
E-mail addresses: mark.s.ibrahim@uvm.edu (M. Ibrahim), chris.danforth@uvm.edu (C.M. Danforth), peter.dodds@uvm.edu (P.S. Dodds).

http://dx.doi.org/10.1016/j.jocs.2016.12.001
1877-7503/© 2016 Published by Elsevier B.V.



Train

From Wikipedia, the free encyclopedia

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

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

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

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

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

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

A BNSF intermodal freight train passes through Wisconsin, United States

Part of a series on
Rail transport

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

First Link Path to Philosophy
beginning with "Train"

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

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; the first link is circled in red. 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. After 11 links, we arrive at "Philosophy." (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

systematically capture the core description of a topic. Within the body text, the first link marks the earliest moment in a topic's introduction where the authors choose to directly reference another article. While many links reference relevant details, the first link is an association within a topic's initial description. While "Amtrak's Acela Express," "steam," and "head-end power" are links detailing particulars, the first link, "rail transport," is the topic the authors associate with "Train" in the introduction (Fig. 1). "Banana" has a first link to "fruit," "Bob Dylan" has a first link to "Blowing in the Wind," and "Physics" has a first link to "natural science." Collectively, first links provide a pragmatic and interpretable means to connect each article to another.

By following the first hyperlink contained in all articles of the English edition of Wikipedia as of November 2014, we form a directed network: *Wikipedia's First Link Network* (FLN). Inspired by the claim that the majority of first links lead to "Philosophy"—popularized by an xkcd comic and subsequently discussed in blog posts [19–23]—we holistically study how the FLN yields a flow of connections. For methodological details, see Appendix A.

The FLN is a wealth of relations among inventions, places, figures, objects, and events across space and time. "Train" for example, links to a parent node, "Rail Transport," while many child nodes such as "Steel" and "Horsepower" link to "Train." As shown in Fig. 1, the path starting at "Train" contains "Goods," "Economics," "Social Science," leading ultimately to "Philosophy". Unlike previous taxonomies created by individuals [24–27], the relations in the FLN emerge without a centralized effort as the aggregate of each article's authors choice of first link.

Our goal is to study the structure of the FLN for insight into how the information on Wikipedia is organized and related. Informed by river network metrics, we consider the FLN as a flow. We quantify the accumulation of first links around articles and develop a new method, *traversal funnels*, to measure 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.

2. Traversing the First Link Network

An essential feature of a directed network's structure is the degree distribution [28]. The degree distribution has been used to study many phenomena from disease outbreak [29] to the dynamics of social networks [30]. The in-degree distribution in the FLN describes how many first links point to a particular article. Articles with zero in-degree have no references—they are outer leaves in the FLN or are disconnected entirely.

Starting at each article, we construct a path through the FLN, to map the 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. Previous studies have used flow to characterize the structure of river networks [31,32] and describe the organization of blood networks, food systems, and transportation networks [33]. In the same vein, we use flow to measure accumulation and develop a new method for isolating influence in a directed network with cycles.

2.1. Traversal visits

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 a page is revisited or is invalid—defined here to mean pointing to a page 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 define the number of *traversal visits* of an article to be the number of references flowing to the ideas in the article—equivalent to drainage 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 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 encoding each path out of the more than googol possible paths ($\simeq 4.7 \times 10^{6} \times 2^{4.7\times10^{6}}$) through the FLN.

By measuring the number of first links between two articles, we obtain an additional piece of information we call *path length*, computed by summing columns in the traversal visits matrix. Path length describes how closely related topics are. Although "Train" is related to "Economics" for example, there are several articles bridging the connection: "Train" is more specifically related to