MPr : 901
Does prerequisites for software and for learning mathematics has any similarities/differences in the network properties

Atthaluri shashank
5th Year Int. M.Sc. Mathematics
Department of Mathematics
UM-DAE Centre for Excellence in Basic Sciences
Electronic address:a.shashank@cbs.ac.in

Project Advisor: Prof. Nagarjuna, G.
Homi Bhabha Centre for Science Education
Tata Institute of Fundamental Research
V. N. Purav Marg, Mankhurd, Mumbai 400088, India.
Electronic address:nagarjuna@hbcse.tifr.res.in

November 19, 2019

Abstract

We propose a simple model for the development of teaching/learning pathways for mathematics education. This model assists in visualization of learning sequence given each learning objective, as a node, is linked with its prerequisites by drawing a directed acyclic graph. The method adopted (1) shows the number of layers of a dependency graph (2) illustrates emerging segregation of different domains of knowledge based on semantic proximity. We then indicate how to create a visual network graph given the nodes. To demonstrate this idea, topics on mathematics were taken from Khan Academy as learning objectives and an analogous complete data set of software package as a sample of dependency data based on previous work. We discuss on the usefulness of this idea for mathematics education.

Acknowledgements

Hereby, I wish to thank, all people who in one way or another supported and guided me through my Master Thesis research project.

First, I want to thank Prof Nagarjuna, G. I have learned many things since I became his student. He spends vary much time instructing me how to write a thesis, how to search literature and how to collect data. He supported me in learning of many new things. I want to thank him as well for answering several silly questions I proposed during this project.

Next, I want to thank Prof Sanjay Chanderasekhar giving me a chance for collaborative discussion in his lab and also allowing me to attend the course which he instruct.

Thirdly, I would like to thank all the scientists I have referred to in my Thesis. I literally spent lot of time for reading and understanding these sources. They supported me with well-funded theories and useful insights.

Fourthly, I am very thankful for Surrendra and Soham for helping me in accomodation and internet facilities. I thank Durga prasad, Rajat Jain, Harshit, Megha, Ganesh, Rohit somanchi and Vibhu Vaibhav for fruitful discussions and debates.

I thank all professors and employees of HBCSE of the last few months for their education and support activities.

Finally, I am grateful to my parents, my family and my close friends. Without their support, and their love, I would not be able to complete this project.

Contents

"*We know very little, and yet it is astonishing that we know so much, and still more astonishing that so little knowledge can give us so much power.*"

Bertrand Russell

Chapter 1

Introduction

"What I do on a daily basis is figure out how networks affect our lives...and lately I am very interested in how networks affect us as individuals and how they determine our success."

Barabasi

Since its rumbling from the late 20th century (Watts and Strogatz 1998; Barabasi Albert 1999), the concept of networks played a crucial role in understanding complexity [1–3], also become continuously relevant understanding everyday phenomena. Nearly every system that surrounds us, or exists among or inside of us, can be understood as a network, i.e., a discrete structure that consists of nodes (vertices, actors) and the edges (links, relationships) that connect the nodes. Examples of networks are ubiquitous, including the Internet [1], social media [4], financial systems [5], transportation networks [6, 7], semantic [8–10], ecosystems, organizations, friendships [11], schools, classrooms, brains [12], immune systems, and even genes/proteins within a single cell. Knowledge about networks can help us to make sense of those systems, making it a useful literacy for people to be effective and successful in this increasingly complex, interconnected world of the 21st century. Network science (Barabasi 2013), an interdisciplinary field of scientific research on networks, offers a powerful approach for conceptualizing, developing, and understanding solutions to complex social, health, technological, and environmental problems.

We develop a simple model to construct a semantic network which are generally contextualized. Several models have been developed to investigate the connection of words from a semantic association perspective [10, 13].
Before introducing to any new topic every good author tries to inform about the prior-knowledge required. This is due to the obvious assumption that if the prerequisites are not satisfied adequately the learner may not be able to comprehend the new ideas introduced in the book. This has been the general guiding principle of curriculum de-

sign. This principle also stands out as one of the consensus from the widely accepted constructivism (philosophy of education). Constantly helping and reinforcing the prior knowledge to learn something new is a time-tested ancient wisdom shared among most educationists. Based on this assumption we propose a simple method of processing prerequisites for conceptual structures by employing a conceptual structure itself. Mastering conceptual structures is a skill that requires inter-disciplinary understanding involving certain topics from domains such as logic, linguistics, mathematics, AI, databases, philosophy, computer science etc. We propose a simple method to construct a machine processable semantic network, which may be called a dependency network, that gathers all the concepts, skills as nodes and the relation type "depends on", as their edges. There are several studies on the dependency relation in different contexts. [14, 15] Semantics and logic of dependency are covered in [16]. Keller identified some kinds of dependencies in the context of requirements analysis [15]. The central idea is to create a database of learning objectives, each of which specified with their prerequisites. For example, "acceleration" is a prerequisite for "force", "velocity" a pre-requisite for both, "speed" a pre-requisite for "velocity", and so on. Such data can be used to draw a directed graph using any of the several graph visualization tools. Based on this idea, a pilot portal was created to collaboratively generate valid dependency relations between learning objectives. In this work, it is claimed that:
Since all learnable concepts and skills can find a place in the dependency map, where each node can have a determinate position with longitude and latitude of the map also specifiable, it is clear that a surface map of all knowledge is indeed possible [14]. This is an extention of this work and demonstrates the possibility of creating multiple visualizations of knowledge. The tools and techniques that we use in this work are well known, but the application of them is novel and we propose it could have useful impact in the field of education. Much of this work is an application of graph visualization and analysis of the semantic system in terms of graph theory. We use two kinds of datasets, one from prerequisites of an educational domain and another from software dependencies of an operating system. In the context of education, there are a few projects that already make use of the graph character explicitly, e.g., Expii (http://www.expii.org),Khanacademy (http://www.khanacademy.org), and Atlas of Knowledge (http://atlas.gnowledge.org/). In the case of software dependencies, a popular distribution of GNU/Linux (Debian) (https://www.debian.org/) manages more than 27,000 packages by explicitly defining their dependencies. We use cartography as a metaphor, because the resulting graphs and layouts that we build produce a navigable 'surface' indicating clearly the application. We call the resulting layouts as a map of knowledge (for 2D). The scope of the graph representation in this work pertains only to the semantics (meaning) in the context of learning. In what follows, we present the specific methods and the tools used in processing the above two datasets and generating the data visualizations followed by the discussion of the results. The datasets, interactive and high-resolution visualizations, algorithms are made available.

Chapter 2

Literature

"Access to more information isn't enough- the information needs to be correct, timely, and presented in a manner that enables the reader to learn from it. The current network is full of inaccurate, misleading, and biased information that often crowds out the valid information. People have not learned that "popular" or "available" information is not necessarily valid."

Gene Spafford

2.1 Network Science Vs Graph Theory

In the scientific literature the terms network and graph are used interchangeably:

Network Science	Vs	Graph Theory
Network		Graph
Node		Vertex
Link		Edge

Yet, there is a subtle distinction between the two terminologies: the network, node, link combination often refers to real systems: The WWW is a network of web documents linked by URLs; society is a network of individuals linked by family, friendship or professional ties; the metabolic network is the sum of all chemical reactions that take place in a cell. In contrast, we use the terms graph, vertex, edge when we discuss the mathematical representation of these networks: We talk about the web graph, the social graph (a term made popular by Facebook), or the metabolic graph. Yet, this distinction is rarely made, so these two terminologies are often synonyms of each other [17].

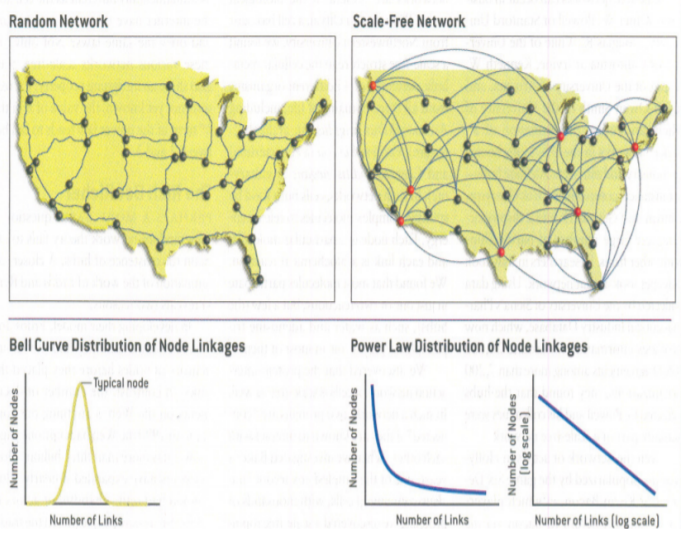
2.2 Degree, In and Out degree, Degree distribution

The degree of a node in a network is the number of connections or edges the node has to other nodes. If a network is directed, meaning that edges point in one direction from

one node to another node, then nodes have two different degrees, the in-degree, which is the number of incoming edges, and the out-degree, which is the number of outgoing edges. The degree distribution P(k) of a network is then defined to be the fraction of nodes in the network with degree k. Thus if there are n nodes in total in a network and n_k of them have degree k, we have P(k) = n_k/n.

2.3 Scale-free Networks

Complex systems can be depicted as networks. Therefore, network analysis is an important tool for the analysis of a complex system. A network is called scale-free if the characteristics of the network are independent of the size of the network, i.e. the number of nodes. That means that when the network grows, the underlying structure remains the same. The underlying structure: A scale-free network is defined by the distribution of the number of edges of the nodes following a so called power law distribution.



2.4 Power Law Distribution

The power-law distribution is that the number of nodes with really high numbers of edges is much higher in the power-law distribution than in the normal distribution. The normal distribution is less frequently observed in networks. This is the case because the distribution of edges in a network is mostly not the result of the sequence of independent quantities. Networks grow over time. Nodes that already have a high number of edges are more likely to see new edges to them established compared with nodes with a lower number of edges. That is the idea of so-called preferential attachment or the rich-get-richer principle.

2.5 Small world

In large scale-free networks the small world principle may often hold. This principle says that any node is on average connected to any other node in a small number of steps, say around 5. The nodes with many connections act as a kind of hub between all the other nodes. [18]

2.6 Dependency Map

An analysis of the prior knowledge of an individual is required before laying down a path of learning for him/her. This path of learning is a dependency map of teaching-learning sequences.

Here, the path of learning is defined as a prerequisite required for the comprehension of a concept. Refer to the flow chart Figure 2.1. [14] The simplest method that was

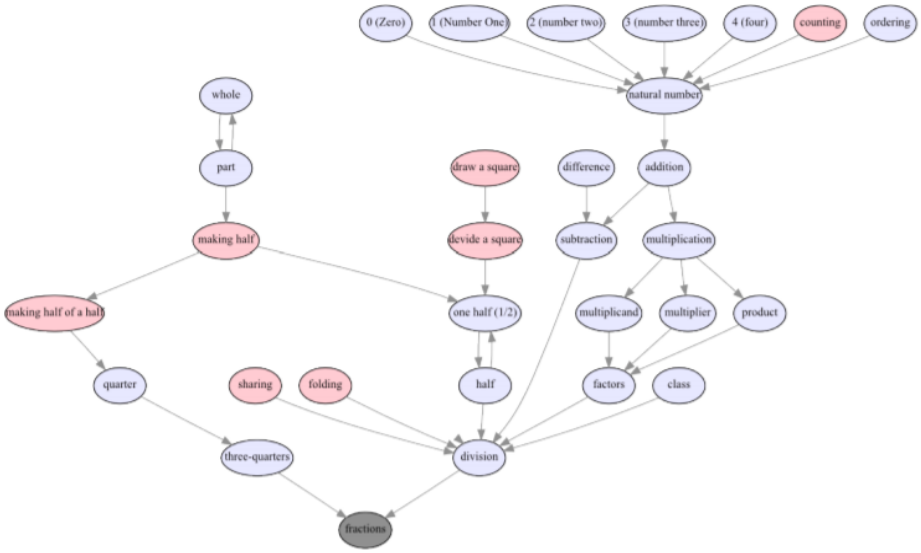


Figure 2.1: An example of a dynamically generated road map. The node for which the sequence is shown is coloured gray. The activities, in contrast to concepts, are represented in pink.

proposed is to gather every assertion of prerequisites from all sources, subjects, and store them in a single large knowledge base. This results in a massive semantic network holding activities and concepts linked by their dependencies. The recursively drawn graph will give us a road map of each learning objective (LO) considering that each node is also an LO.

Here there are two relations between the nodes: "depends on" and "prerequisites for". In flow chart 1 and 2 we define that "multiplication depends on addition" and the

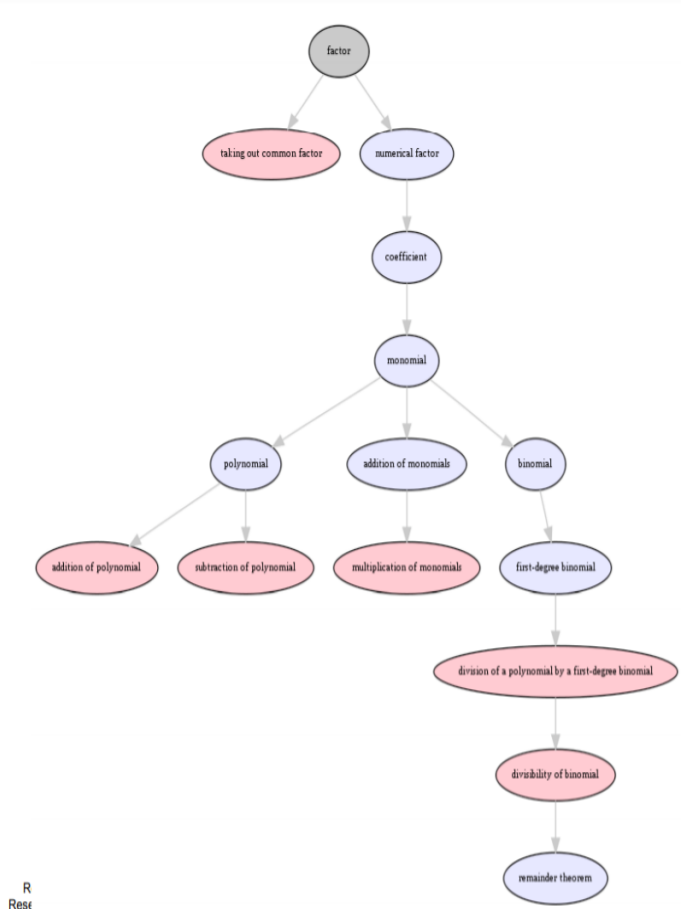


Figure 2.2: The road-ahead map of the concept "factor" dynamically generated based on the data available.

"division is a prerequisite for remainder".

The dependency map is essentially the combination of road map and road-ahead map. Road map has multiple paths that converge to a destination, whereas road-ahead map of any given concept/activity depicts the divergent routes one may explore after reaching a destination.

Chapter 3

Research Methodology

"Although cascading failures may appear random and unpredictable, they follow reproducible laws that can be quantified and even predicted using the tools of network science. First, to avoid damaging cascades, we must understand the structure of the network on which the cascade propagates. Second, we must be able to model the dynamical processes taking place on these networks, like the flow of electricity. Finally, we need to uncover how the interplay between the network structure and dynamics affects the robustness of the whole system."

Barabasi

3.1 The Proposal

Effective teaching and learning require prior knowledge of the learner to be checked so that his/her understanding of all the prerequisite concepts is ensured. Our model aims to detail all the prerequisites that would have to be covered prior to teaching any new topic. This would result in better curriculum design. We used our model on a sample of the Khan academy topics on mathematics where the learning objectives were expressed as nodes that were the concepts. A prerequisite was expressed as a dependency link between two concepts/nodes. For instance, learning addition is a prerequisite to learning multiplication and thus expressed as a dependency link. By gathering all prerequisites and expressing them as dependency links like in the above instance, the graph drawn gives us a road-map of each learning objective.

3.2 Methodology

The cartography of an atlas of knowledge investigated in [14] establishes dependency linkages of any learning outcome (henceforth LO) which depends on pre-requisite LOs. The dependencies in each datasets are processed as a graph. In this model each LO

is defined as a vertex in the graph, while the directed edges indicate the direction of pre-requisite relations amongst the nodes (vertices) in the knowledge network.
The methodology followed is same for both the semantic systems.
"Let the graph G be defined by

$$G = (V, E) \tag{3.2.1}$$

where V as set of nodes and E as set of edges. The order of G, ie the number of vertices, is denoted by $\mid V \mid$ and the number of its edges by $\mid E \mid$.

3.2.1 Layering of Graph (G)

The network is partitioned into discrete layers of nodes [15], where the set of layers L is represented as

$$L = (L_1, L_2...., L_n) \tag{3.2.2}$$

such that each layered set L_i

$$L_i \in V \ for\ all\ i \in n \tag{3.2.3}$$

where all L_i are mutually exclusive subsets of V and

$$\mid V \mid = \sum_{i=1}^{n} \mid L_i \mid \tag{3.2.4}$$

The layer assignment of each node is essentially a maxima function over the preceding layers of each of its prerequisite nodes. The algorithm runs over each directed edge E_i (with Source as S_i and Target as T_i) in the network to designate planes to both the source and the target as shown in Fig. 1. This layering algorithm has a worst-case time complexity of $O(\mid E \mid^2)$ and an average space complexity of $O(\mid V \mid + \mid E \mid)$.

Steps of the Layering Algorithm [19]

1. Assign all vertices to layer 1 such that

$$L(V_1) = L(V_2) = ... = L(V_{|V|}) = 1 \tag{3.2.5}$$

2. Iterate from 1 to $\mid E$
3. Consider each edge E from 1 to $E_{|E|}$
4. Check if $L(V_{Target}) < L(V_{Source}) + 1$
5. If true, assign $L(V_{Target}) = L(V_{Source}) + 1$
6. Quit running iterations if $L(V_j)$ doesn't change for any j in V over continuous iterations.

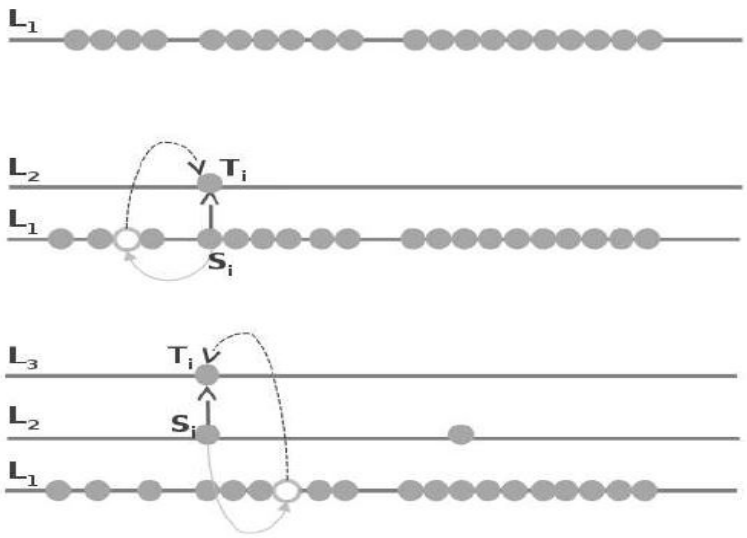


Figure 3.1: Illustration depicting the assignment of layers to each node during execution of the graph layering algorithm.

7. Determine the number of layers n as

$$n = max(L(V_1), L(V_2), ..., L(V_{|V|})) \tag{3.2.6}$$

Chapter 4

Data Analysis and Results

"Tags [distinctive agent features observable by other agents] almost always define the network by delimiting the critical interactions, the major connections. Tags acquire this role because the adaptive processes that modify cas [complex adaptive systems] select for tags that mediate useful interactions and against tags that cause malfunctions. That is, agents with useful tags spread, while agents with malfunctioning tags cease to exist."

John H. Holland

4.1 About data

In the study of software dependencies: the model is explained by the data of packages used while the installation of a software from Debian GNU/Linux distribution [20] where the packages used for the installation are the dependencies, as prerequisites. The mathematical modelling in [20] network is carried out by the collection of data from debian releases, Etch, Lenny and Squeeze. [21].
In the study of Khan academy data dependencies: The data sample taken was concepts covered in topics of Mathematics like Algebra, Trigonometry, and Geometry up to the school level, with the resource being the Khan Academy videos [22] on the concerned topics. The data collection was manually conducted through a thorough investigation of the concept covered in each video. A simple model is proposed gathering prerequisites of all the nodes and check their dependency links. Though nodes were taken as videos, eventually, expression of other terms representing the same concept in each video was checked to properly gather all the prerequisites for each concept.

4.2 Analysis and Results.

4.2.1 Software dependency

The semantic network of nodes in Debian distribution is based on one single principle for all the nodes: Y depends on X; its inverse, X is required for Y. Each package in the Debian system refers to requisite dependencies, with 27,623 such packages and 48,138 dependencies.

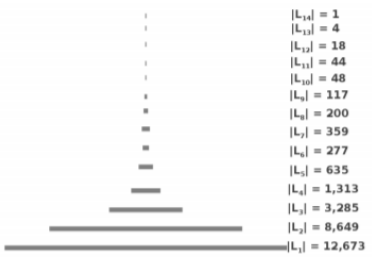


Figure 4.1: Representation of the number of packages (vertices) assigned to each layer in the software dependency dataset (with a depth of 14 layers).

There are 14 graph-layering's of Debian package dependencies. Figure 4.1 explains that the installation of the 13th layer requires prior installation of the preceding 12 layers. Here, the 12 layers thus act as a prerequisite for Layer 13. The characteristics of scale-free network for this system are explained in [20].

A 2 dimensional force-directed algorithm run on the dependency mapping of Debian packages yields Fig. 4.2 which clearly highlights the emergence of several clusters of packages.

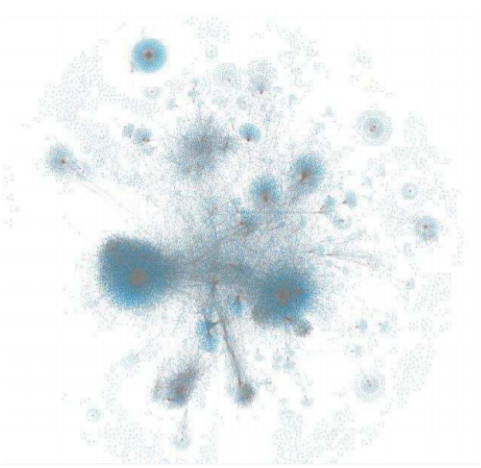


Figure 4.2: Interactive 2-D force directed visualization of software dependency dataset. Clusters are imminently visible on the basis of semantic proximity.

4.2.2 Khan Academy data

The semantic network of nodes in Khan academy data is based on dependency map principle for all the nodes.

Recalling dependency map from Literature [14], here, we have the objectives which depict both convergent and divergent routes. A sequence of learning is thus obtained which cannot be skipped or broken if comprehension of the specified node is desired. As an example, Figure 4.3. where the id's are basically the concepts of videos which tells us that id 3 depends on id 2, implying that we must learn/teach the concept in id 2 before id 3. id 8 depends on id 3 , id 6 and id 7; id 20, id 23, id 28 depends on id 8 and so on. It is insufficient to understand just id 3 for understanding id8 as understanding id 6 and id 7 is also a required.

To understand any concept, it is necessary to understand each prerequisite concept. This flow, as shown in the Fig 4.3 , is a sequence of steps that should be followed. Bridging and walking through all intermediate steps without skipping provides the best teaching-learning sequence.

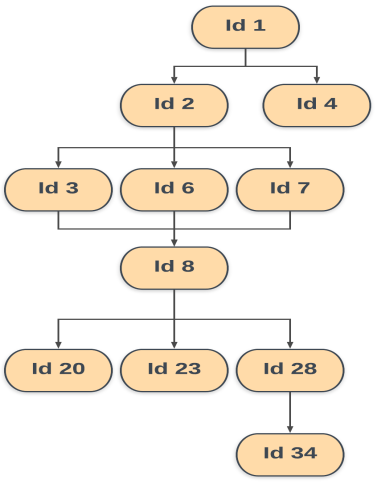


Figure 4.3: Dependency map of the concepts based on small part of Khan academy data.

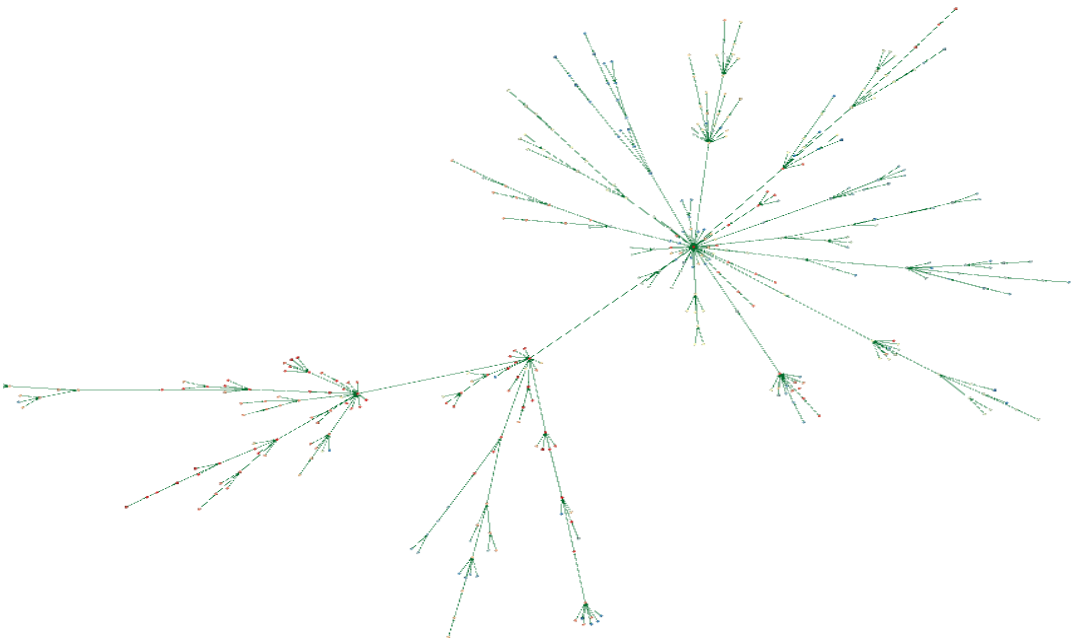
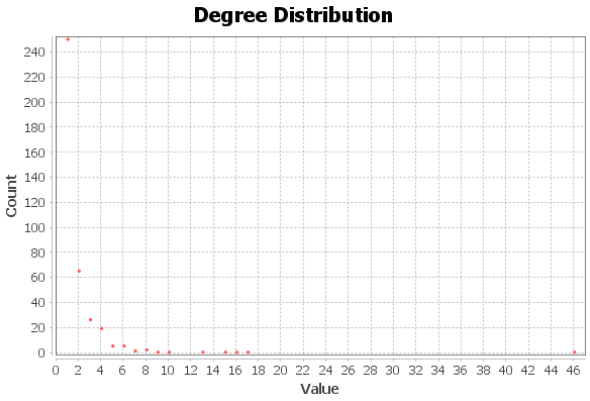


Figure 4.4: Interactive 2-D force directed visualization of Khan Academy Math dataset.

Approximately 400 topics were taken as a sample and data was built with clear prerequisites. The data is collected in a *csv* file [23] with two major inputs which defines the list of nodes(videos) and the prerequisites for certain nodes. By using this data, a directed network graph was drawn using an application Gephi [24] with $yifanhu$ in-built layout algorithm for visualization [25].
The hubs in Figure 4.4 are the evolution from the data which describes the three main topics by their nodes.

The statistical data which we collected using the application Gephi leads to many details of the network as:



Average degree (k). - 0.997

Network diameter is defined as the shortest path between any two nodes in the network - 8

Graph density(D) of a directed network is denoted by $\frac{|E|}{|V|(|V|-1)}$, where $\mid E \mid$ is the number of edges and $\mid V \mid$ is the number of vertices in the graph. Note that the maximum number of edges is $\frac{|V|(|V|-1)}{2}$.- 0.003.

Avg. path length is defined by $l_G = \frac{1}{n(n-1)} \sum_{i \neq j} d(v_i, v_j)$ where n is the number of nodes in G - 2.5.

Avg. Cluster coefficient - 0.

All the above results are obtained manually using certain formulae, but can also be calculated much easier using the Gephi application just by providing the data of source and target.

Chapter 5

Discussion and Future work

"Science is a way of thinking much more than it is a body of knowledge."

Carl Sagan

We observe that these results are important, because visual segregation provides an independent basis for classification of learning objectives into subjects. By using the application, Gephi, 2-D force directed visualization graphs for both the software dependencies and khan academy data were easily constructed. In the case of Khan academy data, the network contains approximately 400 nodes, necessitating magnification to visualize the details. Due to several edges in both the networks, the graph looks quite cluttered, and not very convenient to read. In both the cases the network graph is constructed using the dependencies which leads to graph layering and road map to the objectives. The scale-free features for software dependencies are clearly explained in [20].

The model that is presented here is simple and general. This makes it possible to be employed in all domains of knowledge, leading us to hope that it evolves into a general mapping and sequencing tool for the teaching-learning process. The detailed information provided of all the prerequisites for each concept assists the sequential teaching/learning of all concepts.

Though only concepts covered in Mathematics at the school level were taken as a sample, this model satisfies other streams of science and may be helpful in connecting the entirety of the concepts covered in contemporary sciences as one huge, interconnected network. This is possible because, as the earlier work with this model which is shown on the site illustrates, there is no necessity to redo the entire mapping with increasing data. Instead, it is just a process of adding this data into the existing network. This works for different subjects too, interconnecting them, which satisfies the condition. For instance, the mathematics concepts required to understand concepts in physics is correlated with the same graph, which led us to the probability of interconnecting all of science in a network. The main aim of this project is to achieve network-construction for all subjects, thus streamlining the entire curriculum design process. This will enable

educators to focus completely on conducive learning environments and methodology of teaching.

Chapter 6

Bibliography

[1] S. N. Dorogovtsev and J. F. Mendes, "Evolution of networks," *Advances in physics*, vol. 51, no. 4, pp. 1079–1187, 2002.

[2] M. E. Newman, A.-L. E. Barabási, and D. J. Watts, *The structure and dynamics of networks.* Princeton university press, 2006.

[3] A. Barrat, M. Barthelemy, and A. Vespignani, *Dynamical processes on complex networks.* Cambridge university press, 2008.

[4] G. Nández and Á. Borrego, "Use of social networks for academic purposes: a case study," *The electronic library*, vol. 31, no. 6, pp. 781–791, 2013.

[5] T. Lux and M. Marchesi, "Scaling and criticality in a stochastic multi-agent model of a financial market," *Nature*, vol. 397, no. 6719, p. 498, 1999.

[6] A. Maritan, F. Colaiori, A. Flammini, M. Cieplak, and J. R. Banavar, "Universality classes of optimal channel networks," *Science*, vol. 272, no. 5264, pp. 984–986, 1996.

[7] J. R. Banavar, A. Maritan, and A. Rinaldo, "Size and form in efficient transportation networks," *Nature*, vol. 399, no. 6732, p. 130, 1999.

[8] R. F. I. Cancho and R. V. Solé, "The small world of human language," *Proceedings of the Royal Society of London. Series B: Biological Sciences*, vol. 268, no. 1482, pp. 2261–2265, 2001.

[9] S. M. Caldeira, T. P. Lobao, R. F. S. Andrade, A. Neme, and J. V. Miranda, "The network of concepts in written texts," *The European Physical Journal B-Condensed Matter and Complex Systems*, vol. 49, no. 4, pp. 523–529, 2006.

[10] G. M. Teixeira, M. d. S. F. d. Aguiar, C. F. d. Carvalho, D. R. Dantas, M. d. V. Cunha, J. H. M. d. Morais, H. B. d. B. Pereira, and J. G. V. Miranda, "Complex semantic networks," *International Journal of Modern Physics C*, vol. 21, no. 03, pp. 333–347, 2010.

[11] C. C. Foster, A. Rapoport, and C. J. Orwant, "A study of a large sociogram ii. elimination of free parameters," *Behavioral science*, vol. 8, no. 1, pp. 56–65, 1963.

[12] M. Bota, H.-W. Dong, and L. W. Swanson, "From gene networks to brain networks," *Nature neuroscience*, vol. 6, no. 8, p. 795, 2003.

[13] D. L. Nelson, C. L. McEvoy, and T. A. Schreiber, "The university of south florida free association, rhyme, and word fragment norms," *Behavior Research Methods, Instruments, & Computers*, vol. 36, no. 3, pp. 402–407, 2004.

[14] G. Nagarjuna, "Collaborative creation of teaching learning sequences and an atlas of knowledgge," 2009.

[15] K. Sugiyama, S. Tagawa, and M. Toda, "Methods for visual understanding of hierarchical system structures," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 11, no. 2, pp. 109–125, 1981.

[16] E. R. Gansner, S. C. North, and K.-P. Vo, "Technique for drawing directed graphs," Aug. 28 1990. US Patent 4,953,106.

[17] A.-L. Barabási *et al.*, *Network science*. Cambridge university press, 2016.

[18] D. J. Watts and S. H. Strogatz, "Collective dynamics of âsmall-worldânetworks," *nature*, vol. 393, no. 6684, p. 440, 1998.

[19] G. Nagarjuna, M. Kharatmal, and R. Gupta, "Latitude and longitude of a semantic system from a directed graph of dependencies," in *Unpubhlished work*, p. 13.

[20] R. Nair, G. Nagarjuna, and A. K. Ray, "Finite-size effects in the dependency networks of free and open-source software," *arXiv preprint arXiv:0901.4904*, 2009.

[21] "debian release," *https://www.debian.org/releases/*.

[22] "Khan academy mathematics," *https://www.khanacademy.org/math*.

[23] "csv data," *https://meatthaluri.blogspot.com/2019/11/collected-data-for-network-graph.html*.

[24] "gephi application," *https://gephi.org/*.

[25] Y. Hu, "Efficient, high-quality force-directed graph drawing," *Mathematica Journal*, vol. 10, no. 1, pp. 37–71, 2005.