Six Degrees of Wikipedia

Irfan Dwi Kusuma, 13518060
Program Studi Teknik Informatika
Sekolah Teknik Elektro dan Informatika
Institut Teknologi Bandung, Jl. Ganesha 10 Bandung 40132, Indonesia
13518060@std.stei.itb.ac.id

***Abstract*—"Friend of my friend have a friend who know Donald Trump!", those statement maybe true (or maybe not). But how many minimum "friend of my friend" layer should be there so everyone are connected? There is an unproven graph theory about it called "Six Degrees of Separation" which state that anyone is connected to everyone with "six handshakes of friend of friend". But since there are no way to prove this theory entirely, we can use these terms for other linked things, such as how every actor in hollywood are connected to Kevin Bacon with how many steps, or in this case, how connected a wikipedia article to another wikipedia article through connected hyperlink in one pages or another.**

***Keywords*— Graph Theory, Hyperlink, Six Degrees of Separation, Wikipedia.**

I. Introduction

Six degrees of separation is a popular hypothesis about any people are six, or fewer social connected from each other. This hypothesis originally set out by Frigyes Karinthy, a Hungarian author in 1929. since its inception, there is no way to prove this hypothesis because of the sheer amount of human in the world today.

Six Degrees of Separation in graph theory can be expressed with nodes as the people, and edges for interaction between two people. If somebody knows somebody, there will be a connection between these two. And then all the possible connection between people are connected with edges. And then we can count how many link between me and Barrack Obama, for example.

Even the hypothesis still only become a hypothesis and there are no way to prove it, it's still gaining popularity in the pop-culture. Mainly because of "Six Degrees of Kevin Bacon", which count every single hollywood actor and actress relations to Kevin Bacon, a prominent actor. It also creates a "Bacon Number" because of the degrees of separation between two actor and actress to Kevin Bacon. Bacon's Number starts with Kevin Bacon as Number Zero(0), then actor or actress who played together with Kevin get Bacon's Number 1. And then actor and actress who didn't play together with Kevin Bacon, but played together with actor or actress who played with Kevin Bacon (or in this case, actor with Bacon's Number 1) got Bacon's Number 2, and so on and so on until everyone are connected in some way. If there are none connection between them and Kevin Bacon, they are assigned with Bacon's Number of Infinity.

This Six Degrees of separation is can also used in some other way, including webpages, or E-mail. In 2013, Hungarian physicist Albert-László Barabási conducted an experiment and discovered that degrees of separation between any two pages are 19 degrees average[1]. The Nearest way to prove this connection are a research conducted by Facebook using friend-list in their website, and there are 99,91 % of Facebook users are interconnected, and in February 2016, distance between every users averaged at 4.57 [2]

For Degrees of separation in Wikipedia itself, started at late February 2018 with website of https://www.sixdegreesofwikipedia.com published by Jacob Wenger. The website takes two Wikipedia Articles and then find the appropriate hyperlink that interconnect those two webpages as short as possible (at the least amount of clicks). From these webpages, only 1,417 percent webpages that interconnected with six or more degrees of separation. The search with no connection are also small, only about 1.07 percent. These are caused by dead links and dead end pages.

II. Theories

A. Graph Introduction

A simple graph G consists of a group of nodes of V, and a set E of two-element subsets of V called edges [3]. For example in Figure 1 there are 9 nodes {a,b,c,d,e,f,g,h,i}, and 8 edges {{a,b}{a,c}{b,d}{c,d}{c,e}{e,f}{e,g} and {I,h}}.

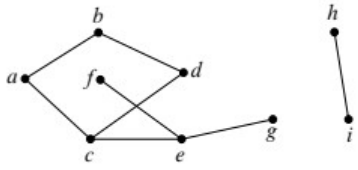


Figure 1 : An example of a graph with 9 nodes and 8 edges. [3]

Two nodes are called adjacent if they joined by an edges, and the edges are incident to the nodes joined. Number of edges between an nodes and other nodes are called the degrees of the nodes.

For Example in Figure one we can se a is adjacent to b and b is adjacent to d. and edges {a,c} connect node a and node c. and we can see the degrees of a is two since it's connected with node b and node c {a,b}{a,c}. This also happened to other nodes and with it's own degrees, such as node I who only have one connected nodes (h), and g with e as the only nodes that connect to it, making it having degrees of one. In simple graph

there is also possibilities that a nodes having degree of 0, in which case the nodes has no adjacent nodes that connected to it. Or even a graph have no edges at all, making every degrees of nodes zero.

In simple graphs, there are no self-loops {a,a} since edge is defined to be a set of two edges. And in simple graphs, there are no backward edges {a,b} and {b,a}, since simple graph does not contain directed edges.

B. Some Common Graphs

Some Graph have their own unique name. This are because there are some graph that have special properties, such as how one nodes connected to another, how many connection in between those nodes and so on. One special names for these special graph is a complete graph. Complete graph are graph which every single nodes in it are connected with each other nodes in the graph for a total number of edges is n(n-1)/2 with n as the number of nodes. These graph are shown in figure 2. If there are complete graph, there are also an empty graph, a graph which every single nodes didn't connect to any other nodes in the graph. Or everything has degrees of zero. An Empty graph are shown in figure 3.

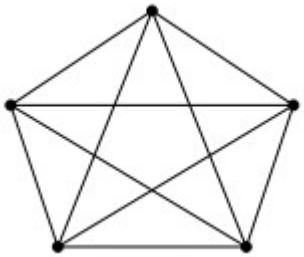


Figure 2 : Complete Graph [3]

Figure 3 : Empty Graph [3]

There are also graph that containing n-1 edges in sequence , such as if the edges are {x1,x2,…,xn} and the edges are {{x1,x2}{x2,x3}….{xn-1,xn}}, the graph will be called a line graph. This because the graph making a line from one point to another. And if the line graph are added and edges of {xn,x1}, it will become an cycle graph, since it's making a cycle from the graph. Line graph and cycle graph are shown at figure 4 and 5 retrospectively.

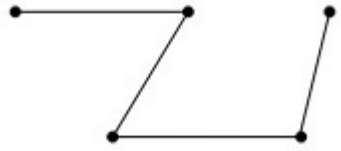


Figure 4 : Line Graph [3]

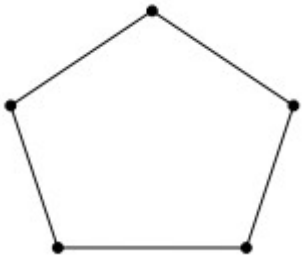


Figure 5 : 5-node cycle graph. [3]

D. Path and Walks.

A walk in a graph is a sequence of nodes and edges such that it's creat a line graph in sub graph itself. For example in figure 1 we can do a walk from node a to f through node c and e. so therefore the walk are using node {a,c,e,f} and edges of {{a,c} {c.e}{e,f}}. These walks start from a and ended in f.

The length of a walk or a path is how many times these walks travel through edges. For example in that walk from a to f have path of four edges. In every walk, there will always be a path.[3]

In a pair of nodes in a graph that are connected by a walk of distance k, it's often many walks that can be used for it. Some example for the walks in figure 1 from node a to node d we can use a,b,d or a,c,d. These are two same distance walks toward the same final destination and also same start point. [3]

E. Directed Graph

Directed Graph, just like undirected graph are graph that also consist of group of nodes and a set of edges. The differences are in the set of edges of directed graph, are specified by an ordered pair of nodes. This makes a difference between {a,b} and {b,a} which in undirected graph are the same. A directed graph are called simple directed graph if there are no loops {a,a} and no multiple edges in the graph. [3]

Directed graphs are commonly found in the application of relationship if the edge is 1-way or asymmetric such as 1-way street, one person likes another but the feeling is not necessarily reciprocated, and sequence of a job which cannot be worked out if the last job has not completed yet. [3]

Directed graph also differ in degrees of the nodes. In undirected graph, there is only one variable which is degree, that shows how many connection between in a node and other nodes. In directed graph, degrees are specified with two variables, indegree and outdegree. Indegree are how many nodes that are connected with a nodes that comes from other nodes {_,a}. And outdegrees are how many degrees that come out from a node into other nodes. In figure 5, we can see an example of a directed degrees and we can see the number of indegree of node c are 2, and the outdegree of node c is 1. there are also special naming system for nodes if the node has zero indegree and outdegree. If there is a node which has outdegree of zero, it's called *sink* and of there are zero indegree in a node, it's called *source*. In Figure 6 there are one source node (a) [3]

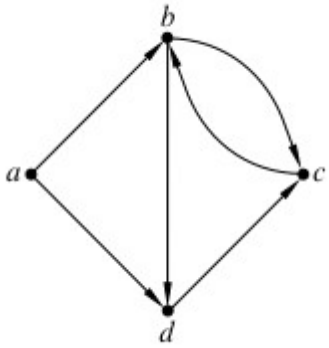


Figure 6 : Directed Graph [3]

In Directed graph, status of one graph being connected are litte bit more complicated than undirected graph. For example, graph in figure 6 is connected or is it not? There are path from a to other nodes, so it's connected. But there are no path into a from other node (b,c,d) so it's not connected? Why there is a contradiction in here? Therefore we cannot do those things. We need to come with a notion of connectivity. Graph theorists has come to a conclusion of creating notion of strong connectivity and weak connectivity. [3]

A Directed graph is called graph with strong connectivity if every pair of nodes has a directed path from one another. For example in figure 6, it's not a strong connectivity graph, since node a has no direct path from node b to node a. we can change it into stronrly connected graph if we remove node a, so node b,c,d will have their own graph which strongly connected.[3]

In Directed graph, walks, path and cycles are similar to the undirected graph except for the direction of the edge must be consistent with how the walk is done. For example walk from c to d in figure 6 is traversed with {c,b}{b,d}. Even though there is a {d,e} walks, there is no {c,d} therefore we must go to node first then go to node d since this is a connected graph.

III. Six Degrees of Wikipedia

Six degrees of Wikipedia is a website which quoted from the website "allows you to find the shortest hyperlinked paths between any two pages on the world's largest free online encyclopedia"[4]. This website contains two box and a result box to shows how the graph of connected articles.

Picture of Six Degrees of Wikipedia Website connecting Spud Gun to Mahatma Gandhi [5]

We can also looking for the individual path of it if we scroll down.

This Website is using data from Wikimedia, which then creates a gzipped SQL dumps of the English language Wikipedia database twice monthly. The Six Degrees of Wikipedia SQLite database is built by downloading, trimming, and parsing the following three SQL tables:

1. page which containing the ID and name (among other things) for all pages.
2. pagelink which containing the source and target pages all links.
3. redirect which Contains the source and target pages for all redirects.

Six Degrees of Wikipedia are only using Wikipedia webpages that are main pages, therefore, no Talk pages or User Pages. In wikipedia, it's a page that belong to namespace 0.

The Database in Six Degrees of Wikipedia is a single SQLite file containing the following three tables:

1. Pages, which contains Page information for all pages.
   I. id – Page ID
   II. title – Page Title
   III. is_redirect – if a page is connected to a page.
2. links, which contains Outgoing and incoming links for each non-redirect page.
   id - The page ID of the source page, the page that contains the link.
   outgoing_links_count - The number of pages to which this page links to.
   incoming_links_count - The number of pages which link to this page.
   outgoing_links - A |-separated list of page IDs to which this page links.
   incoming_links - A |-separated list of page IDs which link to this page.
3. redirects - Source and target page IDs for all redirects.
   source_id - The page ID of the source page, the page that redirects to another page.
   target_id - The page ID of the target page, to which the redirect page redirect

The Owner of the website also keep the search result in a Separate SQLite Databasewhich contains a single searches table with the following schema:

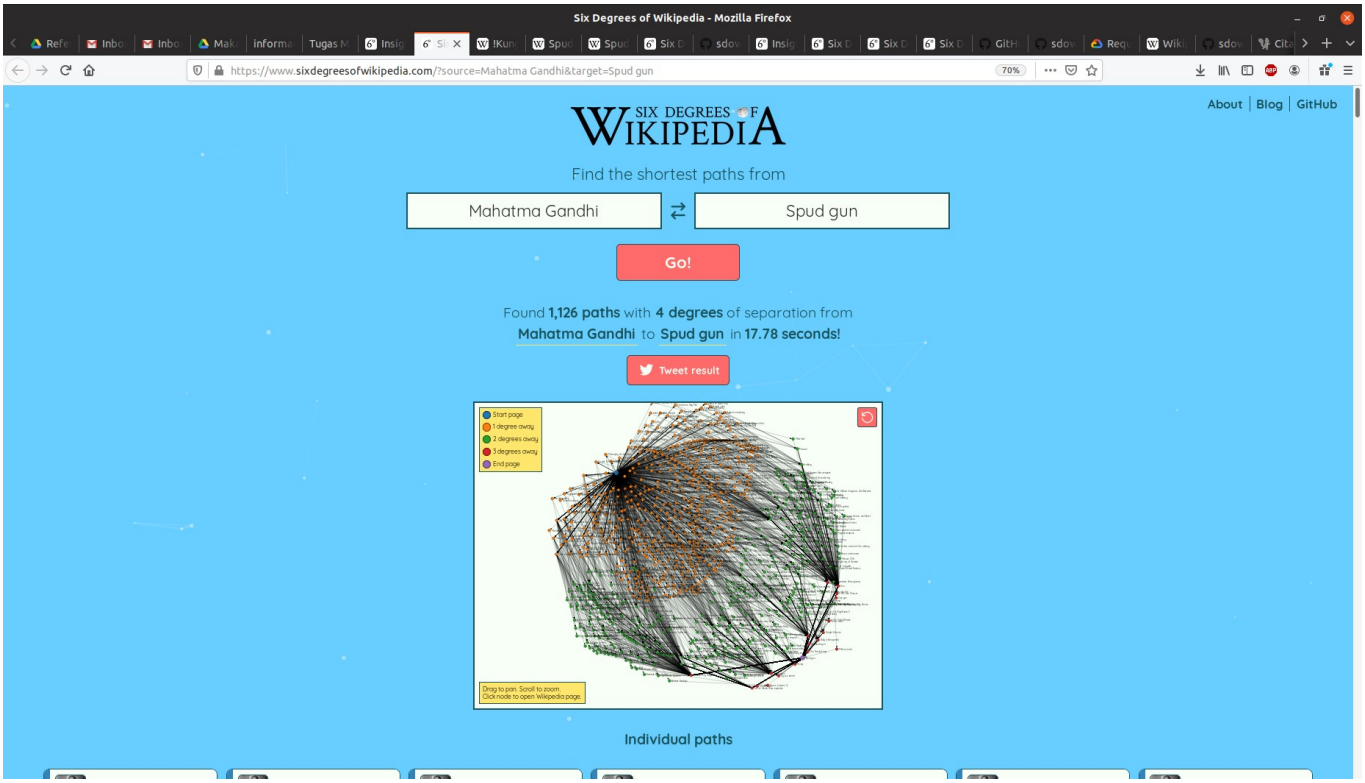
1. source_id - The page ID of the source page at which to start the search.
2. target_id - The page ID of the target page at which to end the search.
3. duration - How long the search took, in seconds.
4. degrees_count - The number of degrees between the source and target pages.
5. paths_count - The number of paths found between the source and target pages.
6. paths - Stringified JSON representation of the paths of page IDs between the source and target pages.
7. t - Timestamp when the search finished.

Search results are kept in a separate SQLite file to avoid locking the main sdow.sqlite database as well as to make it easy to update the sdow.sqlite database to a more recent Wikipedia dump.

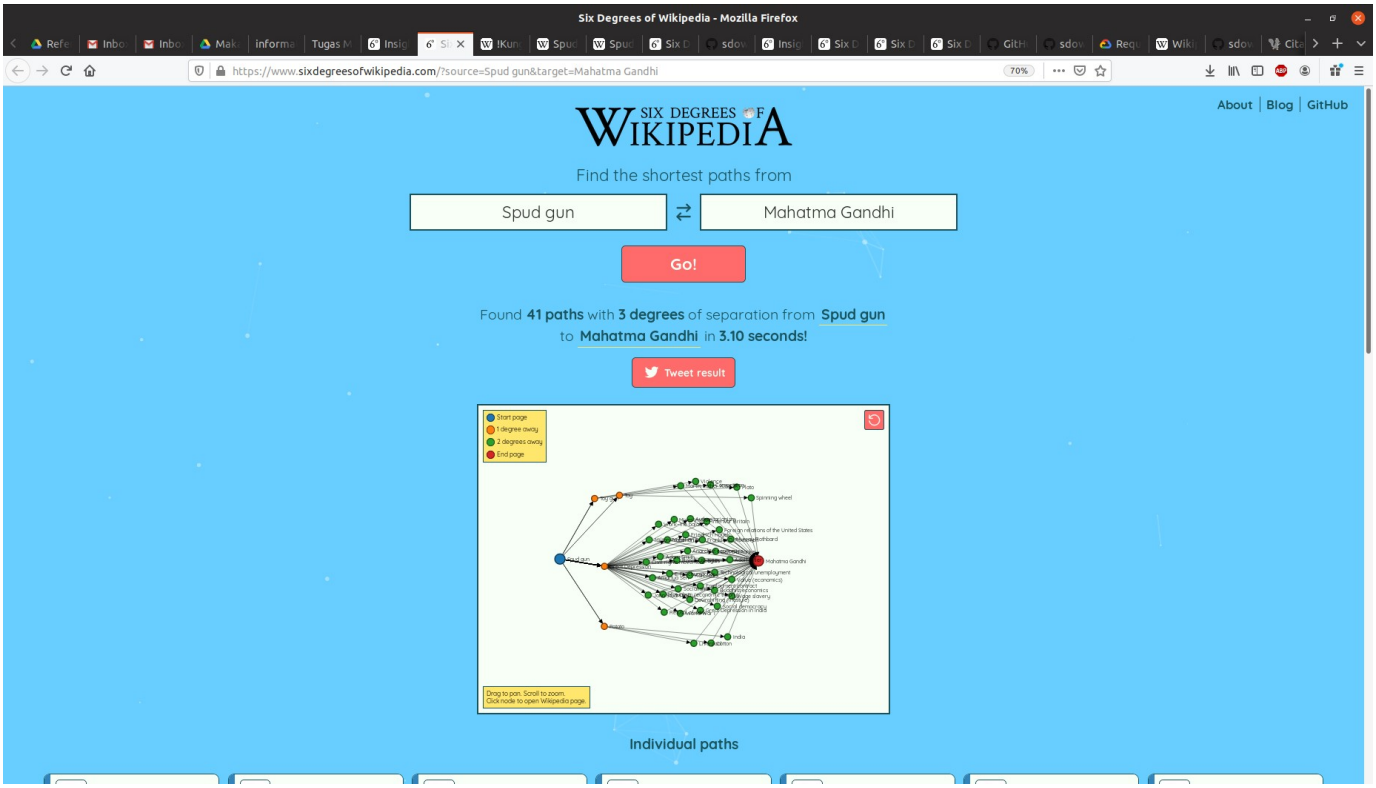
These Search Result are kept for the owner of the website and then analyzed. With the data analyzed, we can see a very interesting result. Such as the most searched result, (which is Adolf Hitler) and other things such as how many Average Degrees of Separation, and other things.

Six Degrees of Wikipedia is a directed graph, since a directed graph is not always strongly connected, there are connection which there are one way, but no way back ({a,b} but no {b.a}). This is because every pages are unique on it's own way and so has their own hyperlink.

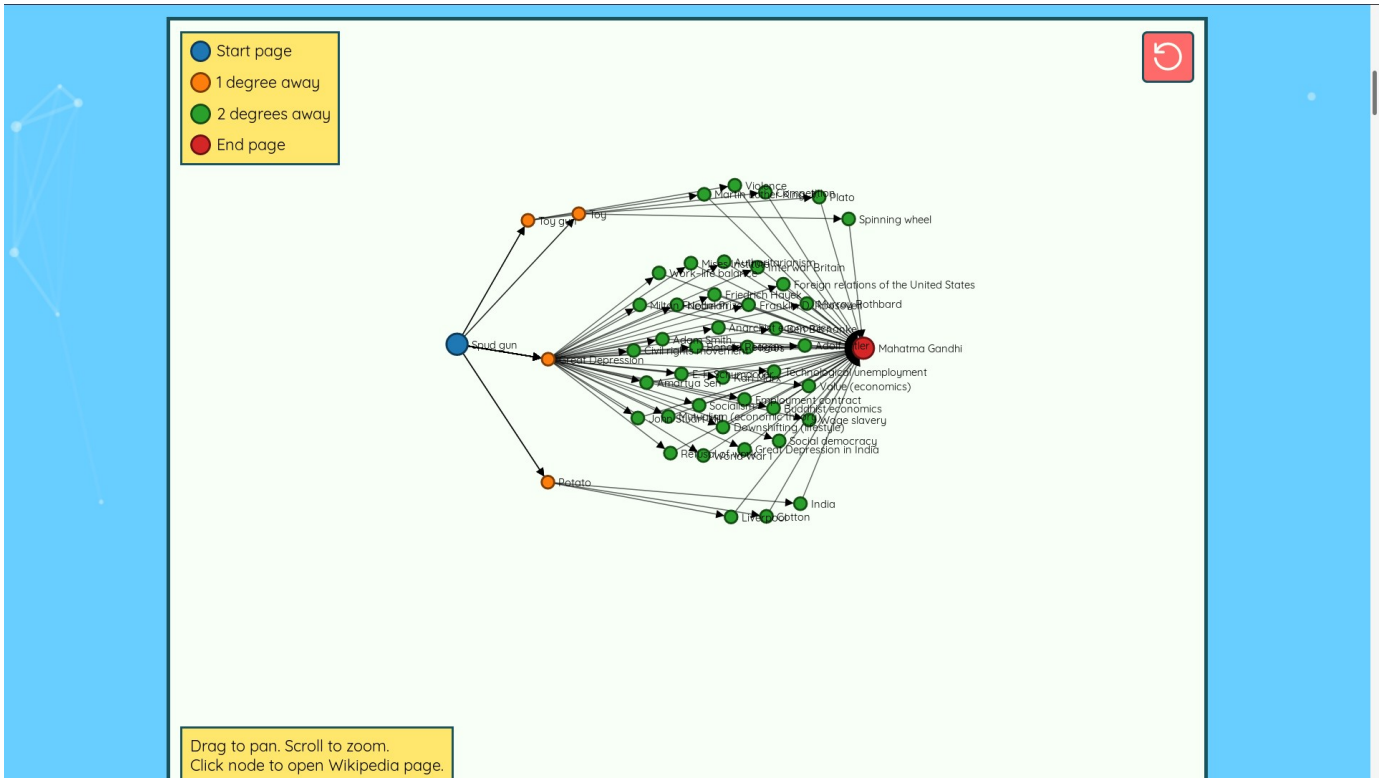
There are also different degrees from one pages to another and backward. Such as Spud gun (a toy) to Mahatma Gandhi need three degrees so it can arrive, but Mahatma Gandhi to Spud gun need four degrees.



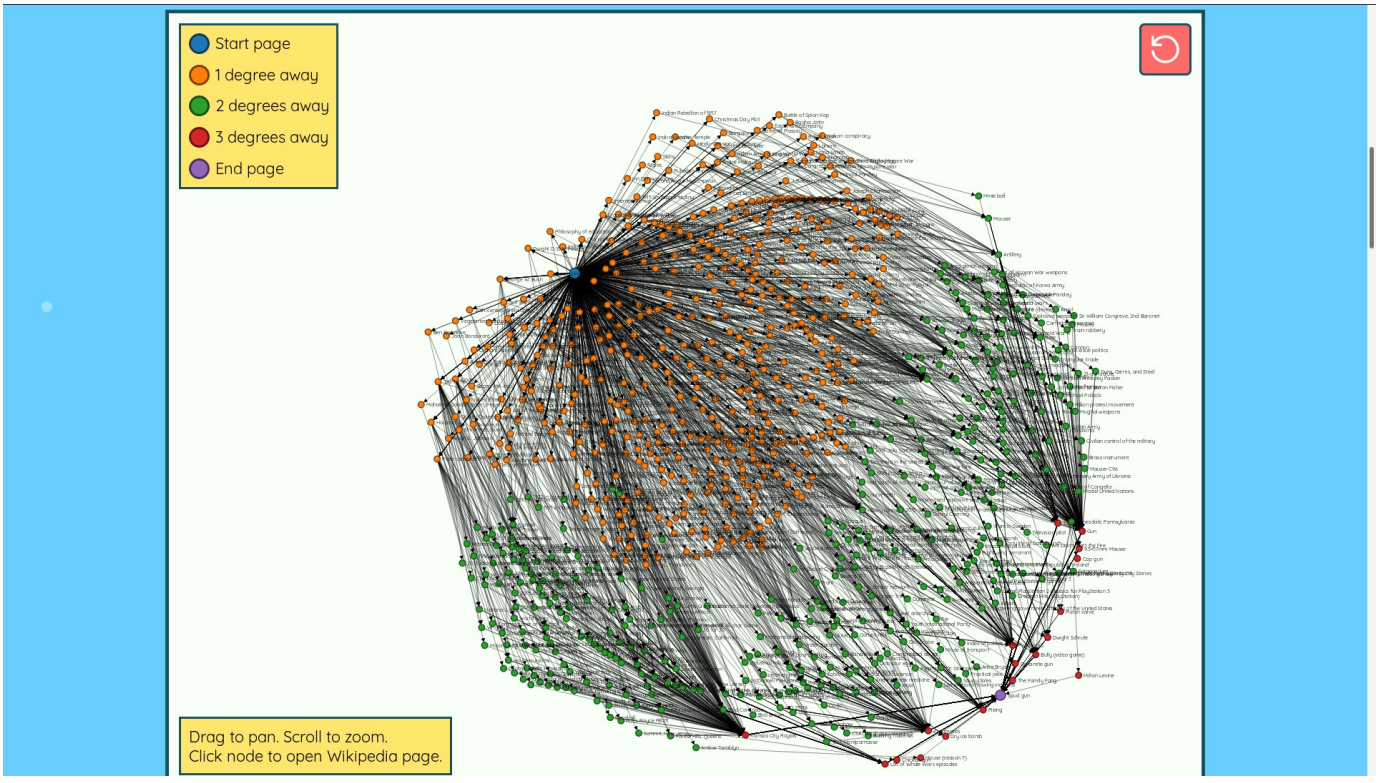
Mahatma Gandhi to Spud Gun Need Four Degrees of Separation



Spud Gun to Mahatma Gandhi on the Other Hand, only need three



Graph From Spud Gun To Mahatma Gandhi



Graph From Mahatma Gandhi to Sput Gun

V. Conclusion

Six Degrees of Wikipedia is an application of graph theory, especially Directed Graph and Shortest Path between two nodes of a graph. We can see that with enough connection, everything can be traced for at few steps Six Degrees of Wikipedia Using SQLite to make the database system, and using React to do the front-end.

VII. Acknowledgment

The preferred spelling of the word “acknowledgment” in American English is without an “e” after the “g.” Use the singular heading even if you have many acknowledgments. Avoid expressions such as “One of us (S.B.A.) would like to thank ... .” Instead, write “F. A. Author thanks ... .” Sponsor and financial support acknowledgments are placed in the unnumbered footnote on the first page.

References

[1] Ionescu, D., & Ionescu, D. (2013, February 19). Any two Web pages are separated by just 19 clicks, study finds. Retrieved December 5, 2019, from https://www.pcworld.com/article/2028714/any-two-web-pages-are-separated-by-just-19-clicks-study-finds.html.

[2] Bhagat, S., Burke, M., Diuk, C., & Filiz, I. O. (2017, January 14). Three and a half degrees of separation. Retrieved December 5, 2019, from https://research.fb.com/three-and-a-half-degrees-of-separation/.

[3] Eric Lehman. 2017. *Mathematics for Computer Science*. Samurai Media Limited, , United Kingdom.

[4] Wenger, J. (n.d.). Six Degrees of Wikipedia. Retrieved December 6, 2019, from https://www.sixdegreesofwikipedia.com/blog/search-results-analysis.

[5] Wenger, J. (n.d.). Six Degrees of Wikipedia. Retrieved December 6, 2019, from https://www.sixdegreesofwikipedia.com/?source=Spud gun&target=Mahatma Gandhi.

Pernyataan

Dengan ini saya menyatakan bahwa makalah yang saya tulis ini adalah tulisan saya sendiri, bukan saduran, atau terjemahan dari makalah orang lain, dan bukan plagiasi.

Bandung, 6 Desember 2019

Irfan Dwi Kusuma
13518060