WikiLinks: Visualizing degrees of separation using Wikipedia articles

Matthew Pick
Stanford University
mpick@stanford.edu

Ashley Ngu
Stanford University
ashngu@stanford.edu

ABSTRACT
Articles about people on Wikipedia are generally limited to people who are considered "notable" by Wikipedia's standards. People with Wikipedia articles are generally public figures or people who have done something "significant, interesting, or unusual enough to deserve attention or to be recorded."[6] Connections between people have been demonstrated using social network data, but these visualizations not optimized for demonstrating and contextualizing connections between notable people. This paper introduces WikiLinks, an interactive tool that allows users to explore collapsible radial Reingold-Tilford trees visualizing the social circles of a person on Wikipedia. This paper describes the design and implementation of WikiLinks and discusses how interacting with these trees on a macro category level and a micro individual level allow a user to gain understanding of the kinds of people a person is associated with.

Author Keywords
Visualization, graphs, data mining, Wikipedia, social networks

ACM Classification Keywords
H.5.m Information interfaces and presentation (e.g., HCI): Miscellaneous.

INTRODUCTION
How can degrees of separation between people across the world be visualized using Wikipedia's vast network of articles?

How can Wikipedia's crowdsourced system of article links be traversed to build an understanding of personal networks, including both direct personal connections and indirect connections.

We built a tool that visualizes the social circles of people using Wikipedia articles. WikiLinks takes in a central person and crawls through their Wikipedia page looking for links to other Wikipedia articles about people. We show summaries about each person, and to contextualize personal connections, we show the sentence containing the article links. A list of the most frequent categories are determined using a person's Wikipedia categories and nodes are colored accordingly.

WikiLinks mines Wikipedia for information about people's social connections and displays that information. For instance, Barack Obama can be linked to Ted Kennedy by way of John Kerry who preceded him as the Democratic presidential nominee. John Kerry served with Ted Kennedy as a Massachusetts Senator.

WikiLinks visualizes the general categories of people who make up a person's social circles, as well as the individual connections. Using this interactive tool enables people to quickly grasp the quantity and context of connections a person has. We use a collapsible radial Reingold-Tilford tree to display these connections, and hovercards to display additional details.

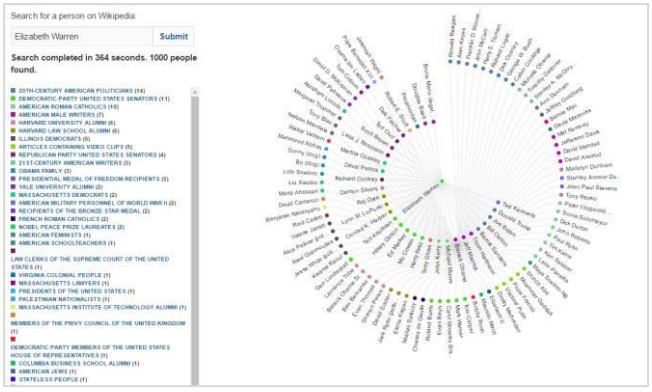

Figure 1, Resulting tree for Elizabeth Warren with Barack Obama's depth 1 node expanded. Warren is the senior U.S. senator from Massachusetts.

RELATED WORK

Initial inspiration for WikiLinks came from the desire to understand the social circles of notable people, as well as Stanley Milgram's work regarding social capital and interconnectedness. In *The Small World Problem*, he posited that people in the US are connected by 3 degrees of separation.[4] Within research fields regarding social influence and connection, Stanley Milgram's work is oft cited. Milgram asked well-connected people to route postcards to a specific random recipient solely through direct acquaintances. If a person personally knew the recipient, that was the end of the chain. If the person did not know the recipient, they were to forward the card to

someone they knew who would be most likely to know the final recipient. The experiment used postal mail and pen and paper, which led to a reliance on geographic proximity and human routing. In our project, we use crowdsourced Wikipedia data to determine links between people.

In order to visualize the social circles of people, we decided to use a Reingold-Tilford d3.js graph to clearly demonstrate the link depth (i.e. degrees of separation between people) using layered graphs. Rather that provide a traditional horizontal nodes and edges plot, we opt for a radial version of the Reingold-Tilford graph to emphasize the relation to social circles. In their paper, Reingold and Tilford propose an algorithm which produces "tidier drawings of trees." This algorithm emphasizes using as little space as possible while satisfying aesthetic rules that aid legibility.[5]

Other related work includes prior social network visualizations that explore social connectivity. *Vizter* from Heer and Boyd explores the topic of social connectivity using social network data from Friendster and also proposes some solutions to navigated large graph structures.[2] Vizster includes features such as the ability to hide or show egocentric networks of each ndoe, and highlighting of groups of connected nodes using color to denote network distance. Vizster utilizes a force-directed algorithm which aids the grouping of users into communities based on connectivity. We learned much from the design choices of Vizster regarding hover information, animations, and node interaction.

In 2012, Daraghmi and Ming Yuan used “graph theory to re-verify the small world theory in an online social network.” [1] This group of researchers concluded that there are four degrees of separation in the world using a dataset from Facebook. Our Wikipedia dataset differs from the Facebook dataset in that all links are based on crowdsourcing and publicly available facts. Our dataset emphasizes any notable relational link, not limited to confirmed friendships. For instance, two people who are not on friendly terms would not be linked on Facebook, but could certainly be linked on Wikipedia. Additionally, important influencers or similar people are sometimes linked together on Wikipedia, even if they do not know each other personally.

Although our project draws inspiration from the above previous work, we differ from previous research in this field which mostly deals with direct personal connections between people. Although the Vizster application illustrates digital connections on social networks, these relationships are still rooted in the real personal connections. Conversely, our project also shows the connection between people who may not personally know each other but whom have some manner of relation. Wikipedia links between people can carry significant meaning and help build an understanding of a person by understanding to whom they connect. Speaking metaphorically, if we imagine that each person is a puzzle piece, we can piece together the shape of a single person by fitting together its surrounding puzzle pieces. For example, viewing the people links on George R. R. Martin's page shows connections to science fiction/fantasy writers such as J R. R. Tolkien and H. P. Lovecraft but he also links to some unexpected people, such as William Shakespeare and Sophocles due to similarities in tragic writing style. Observing this knowledge in a visualization can be extremely useful in drawing conclusions about Martin and his style of writing since the connections on Wikipedia are not limited to people who are living, unlike Facebook, and carry deeper meaning as each link was carefully selected on the page to describe Martin and his works.

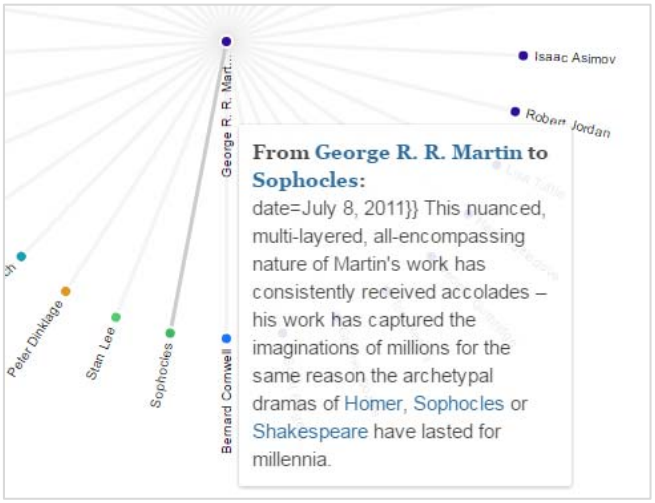

Figure 2, Sentence connection from George R. R. Martin’s page to Sophocles, upon hover over edge. Martin is an American novelist.

METHODS

In order to address our initial question of visualizing links between Wikipedia pages and drawing inspiration from the Wikipedia link hopping game, we initially created a prototype tool whereby users could enter a start and end node and view all possible paths between them. Although this concept revealed interesting paths between seemingly disparate pages (such as Barack Obama → Columbia → Tarantulas), it did not provide particularly useful information about any of the pages in the path. This prototype demonstrated that the sheer volume of links on a Wikipedia page means that very little can be learned without a more specific approach.

Thus, we decided to narrow our focus to pages about real people. This specialization allows us to draw greater insights from the page data in our visualization. Our approach allows users to enter the name of a person as the central node and crawls through their Wikipedia page looking for links to other articles about people. See

http://stanford.edu/~mpick/WikiLinks/ for a live interaction version of WikiLinks.

Wikimedia API

We use the Wikimedia API to retrieve the HTML of Wikipedia pages. Although the API allows users to request the links on a page, it only displays a limited number of links and requires the application to re-query multiple times in order to obtain the entirety of the links. Thus, we are unable to use this feature to obtain links as continuously querying the API significant slows down our application. Instead, we request the page's HTML from the API and use regular expressions to obtain the links. In order to further limit the number of web requests, we limit our API calls to request 50 pages at a time, thereby maximizing the number of pages the API will return in a single call. In order to adhere to the API's terms of use which require calls to be serialized rather than parallelized, we implement a Javascript Queue which manages all of the pages for which we will request data and ensures only one request is made at a time.

Determining Personhood

In order to limit our page link crawl to Wikipedia articles about individual people, we had to determine whether an article was about a person or not. At first, we tried to use Google's Knowledge Graph Search API, which would return whether an entity was a "Person." However, the query limitations were restrictive and made our visualization slow. We then decided to parse the page data of the Wikipedia articles returned from the MediaWiki API, as we already pulled and relied on this data for other WikiLinks features. Articles do not have a direct attribute indicating personhood, but we are able to indirectly infer personhood based on the categories each page belongs to and the Infobox present at the top of most Wikipedia pages.

Most Wikipedia articles about people belong to a "births by year" category such as "1963 births" or "420s BC births." Additionally, many people who are currently alive belong to the "Living people" category. By searching for matches with these categories, we are able to indirectly determine whether an article is about a person. This is critical to our visualization because Wikipedia includes a wide range of articles not limited to people, and our visualization needs to focus solely on real people. Although categories allow us to accurately classify most people, there are certain cases where a person will not have a "Living people" or "birth" category but will instead list their birthdate in the Infobox. Thus we also use regular expressions to search for the presence of a birth date in the Infobox to indicate personhood.

Building the Radial Reingold-Tilford Tree

When a user enters the name of the person for their central node, WikiLinks begins building the tree. The tree is constructed through a breath first search which adds nodes to each ring of the tree and then iterates over each node to build the next ring. The tree is limited to a depth of 2 (ending with the children of the links reachable from the central node's page) because initial testing indicated that people beyond the second depth were mostly irrelevant in their connection to the central node.

We made several design choices during the build phase to emphasize the importance of the tree's structure and category values. First, we choose not to display the names of depth 2 nodes. Although the depth 2 nodes are important in establishing context for the central node, displaying their names during the build phase is distracting and lessens the importance of the depth 1 nodes that are directly located on the central node's page. The depth 2 nodes provide context through their categories rather than deep insight from their actual identities. Second, we choose to collapse a depth 1 node once we finish drawing all of its children during the build phase. Leaving all of the nodes fully expanded would increase visual clutter and distract users from the structure of the tree that is shown as new nodes are added to their depth 1 parents.

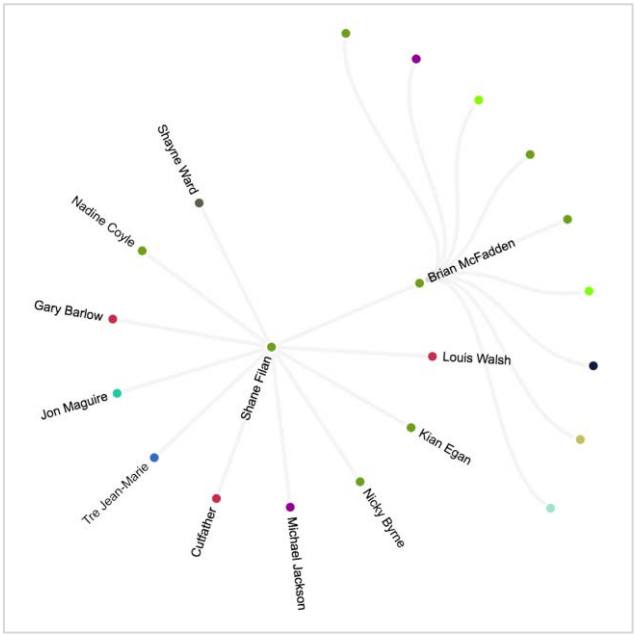

Figure 3, Depth 2 node names are not displayed during the build phase to emphasize focus on tree structure and categorization.

Once the build phase is complete, trees with multiple depth 1 nodes will collapse to only display the central node and its children and the user is able to fully interact with the visualization. If the tree's central node only has one depth 1 child, we expand the tree fully to show the depth 2 nodes and disable expanding/collapsing capabilities. Aside from single child central nodes, users can now explore the collapsed tree by clicking on nodes to expand and collapse their children.

The user's ability to expand and collapse nodes is not enabled until the tree is finished building; however, hover information is still available in the build phase. At any given point while exploring the tree, only one node can be expanded with any open paths collapsing when a user clicks to expand a new node. This design choice was made for users to focus on the current path they are exploring since it is unlikely they would need to compare the children of two depth 1 nodes. Although depth 2 nodes collapse, the central node's children are always displayed as they are the most important connections in the tree.

Categories

Each person is categorized using Wikipedia's category system. WikiLinks assigns each person to one of its categories. The assigned category is the category which most frequently occurs in the set of all visible nodes in the visualization. Nodes are colored according to their category color, which are randomly assigned to each color.

Additionally, the people nodes are reorganized as the tree builds to cluster people of the same category. This makes the tree easier to understand as there is a clear method of organization amongst the categories and users can visually note which categories are the most frequently occurring through larger clusters.

Upon hovering over a category in the legend, the people nodes in that category will be highlighted and bolded. Clicking the category opens up the Wikipedia category page in a new tab.

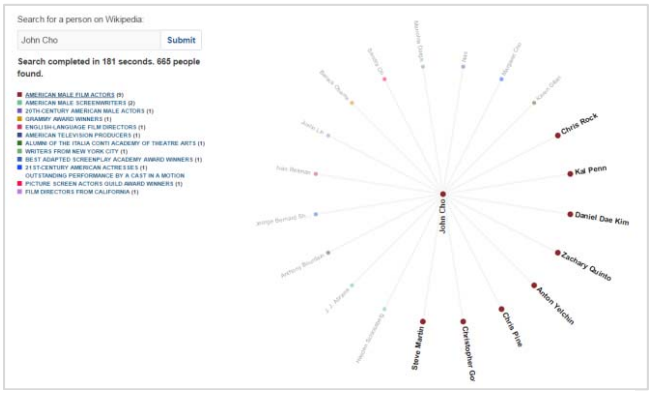

Figure 4, Hovering over the "American Male Film Actors" category with John Cho's tree. Cho is an American actor and musician.

Context Features: Summaries and Link Sentences

While the categories provide general information about the people associated with the central person, WikiLinks also provides information about specific individuals.

In order to contextualize each person and their individual connections, we include a summary of each person, taken from the first couple sentences on their Wikipedia page. These summaries are visible upon hovering on a person node.

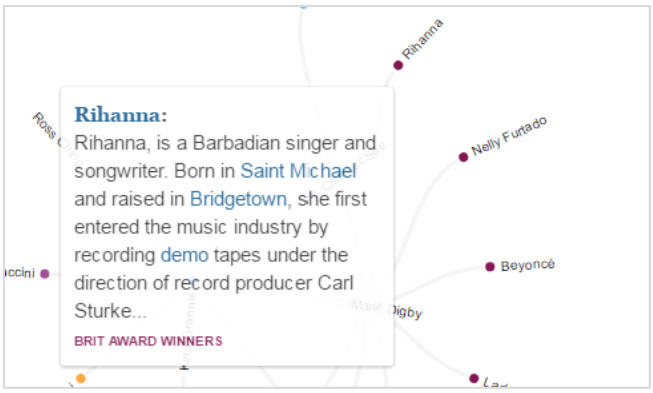

Figure 5, Rihanna's summary, upon hover over node. Rihanna connects to Marie Digby, who connects to Kina Grannis, the central node. Granni is an American guitarist and singer-songwriter.

Upon hovering on connections between people, we also display the sentence in which the page link is found. This contextualizes how one person is affiliated with another. For instance, on Bill Nye's page, Ellen DeGeneres's page link is included in the following sentence, which we display along the edge connection from Nye and DeGeneres:

"Nye's Science Guy personality appears alongside Ellen DeGeneres and Alex Trebek in a video at *Ellen's Energy Adventure*, and attraction that has been playing since 1996 at the Universe of Energy pavilion inside Epcot at Walt Disney World."

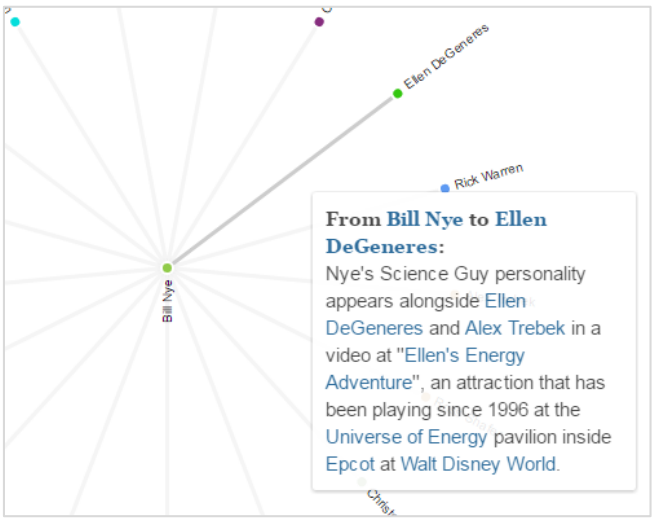

Figure 6, Sentence connection from Bill Nye's page to Ellen DeGeneres's page, upon hover over edge. Nye is an American science educator.

This sentence contextualizes how Bill Nye and Ellen DeGeneres are connected. This feature helps users understand expected and unexpected connections between two people on Wikipedia. Page link connections are

sometimes non-obvious and these sentences provide more details about those types of relationships.

We considered removing these edges because they contribute to significant visual clutter. However, we judged that the contextual value and hover information they provide outweighed the visual cost. As a compromise, we made the edges semi-transparent and opaque upon hover.

RESULTS

WikiLinks set out to help users understand the social circles of notable people. By visualizing the social circles of a single person, grouping people into general categories, and providing specific details about the nature of a connection between two people, WikiLinks is able to accomplish this goal.

Generally speaking, categorizations on WikiLinks are useful for grasping a macro understanding of the types of people a person associates themselves with. For instance, Kina Grannis, an American guitarist and singer-songwriter, is directly connected to people assigned to the categories "American internet celebrities," "BBC Radio 1 presenters," "American female pop singers," "American YouTubers," "American hip hop singers," and "American singer-songwriters." This sort of aggregation and tallying of people into categories aids the user in understanding who Kina Grannis is by painting a picture of who she associates with.

Our categorization of nodes is reliant on Wikipedia categorizations and currently bluntly optimizes for frequency. WikiLinks currently cannot pick the most descriptive category for each person, and doesn't account for relationships between categories. For instance, Elizabeth Warren's tree includes a large number of "20th-century American politicians" as well as "Democratic Party United States Senators." Arguably, those in the latter category are a subset of the former category. Additionally, in Warren's tree, Barack Obama is listed as an "American Male Writer" which is correct, however, is not the best descriptor of him.

Our ability to determine personhood of an article is dependent on the correct categorization of articles. There are some articles who are about people, but are not determined to be people with our personhood verifier because they lack the appropriate categorization or Infobox birth field. Our method of determining personhood results in some false positives and false negatives. This error is more common with individuals who are not high-profile because the more notable a person becomes, the more robust their Wikipedia article becomes.

Tree loading times are fairly consistent, varying slightly based on the response time of the MediaWiki API. Well-connected people on Wikipedia have several hundreds of connections. Given the query limitations of the MediaWiki API, it sometimes takes several minutes for a tree of a person with many page links to other people to load. For instance, the 14th Dalai Lama is extremely well-connected with 1887 nodes and his tree loads in 541 seconds. Bill Nye's tree has 521 nodes and takes 167 seconds to load. Less notable people such as Kina Grannis who has a tree with 81 nodes that load in 22 seconds. We charted a number of examples to demonstrate that there is a linear relation between the number of nodes a tree has and the time the tree takes to load.

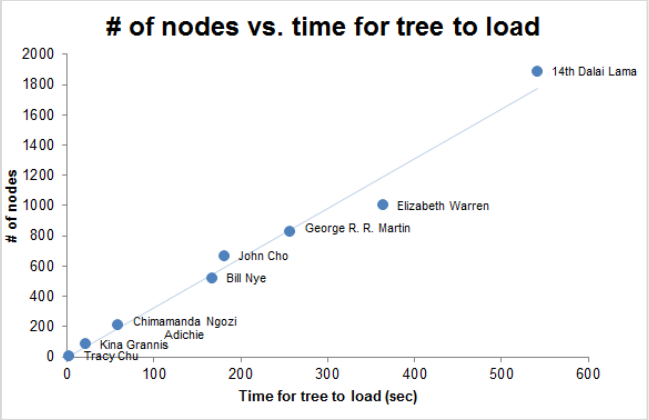

Figure 7, Graph demonstrating linear relation between the number of nodes and time taken to load the full tree.

DISCUSSION

The Wikipedia article network is a unique source of data about the social connections of notable people. By using this data, WikiLinks provides a novel way to understand the a person through the people they are associated with on Wikipedia.

This visualization is a foray into using Wikipedia as a network to learn about people, not just generalizable knowledge. Wikipedia holds advantages over social networks like Facebook or Twitter in that people do not have to explicitly follow or friend others. Information about personal connections between people on Wikipedia is biased towards notable people and towards noteworthy events.

Wikipedia's stated purpose is to act as an Internet encyclopedia, but its structure and content also make it an ideal candidate for analysis as a crowdsourced social network. WikiLinks could be used as a research tool for students, as an alternate method of learning about a notable person. Instead of scanning through the text of Barack Obama's page for names, users can quickly learn about the people he connects to, as well as how he connects to them.

If we visualize Meryl Streep's tree, we are only privy to other people notable enough to have a Wikipedia page and to whom she is associated with publicly. We do not know who Streep's non-notable friends, family, and acquaintances are. This is advantageous for WikiLinks, as when a user is looking into Streep's social circles, it is more important to focus on Streep's public connections rather Streep's private social life.

FUTURE WORK

WikiLinks provides a perspective centered on the circle of one person, but there is further work to be done with more refined clustering using a network emphasis instead of a tree emphasis. Such an extension would aid more complex analysis between nodes, perhaps with bidirectional and loops allowed instead of a unidirectional, acyclic tree. It would also be interesting to use Wikipedia categories to follow a specific group of people over a significant period of time showing generational change. For instance, one could show the networks of American politicians since the Founding Fathers to the modern-day. We hypothesize that one would find compelling connections and overlap between various political periods as notable politicians rose in the ranks, peaked, and then fell out of favor or retired. Additionally, there is bound to be significant mentorship, successorship, colleagueship, and opposing relationships in such a field as politics. WikiLinks would provide an intriguing overview of broad trends, as well as details about individual relations between notable American politicians.

At present, WikiLinks takes one person and visualizes their links, and the links of their links. Instead of stopping at this depth, users could click on a new node to make it the new central node, and continue exploring the Wikipedia network further without having to enter another search term.

Furthermore, we could improve upon the sentence link connection by using NLP to categorize a relationship e.g. family member, friend, colleague, romantic partner, mentor, influencer, successor, or otherwise.

An interesting extension of this system could be to apply WikiLinks with Wikipedias in languages other than English. Wikipedias in other languages have different styles, rules, and structure.

ACKNOWLEDGMENTS

We thank the CS 448B teaching team and our peers, as well as the MediaWiki staff, who provided helpful feedback throughout this project. Additional thanks to all the people who contribute knowledge to Wikipedia.

REFERENCES

1. Daraghmi, E.Y. and Ming Y.S. Using graph theory to re-verify the small world theory in an online social network world. In *iiWAS* 2012.
2. Heer, J. and Boyd, D. Vizster: Visualizing Online Social Networks. In *IEEE Symposium on Information Visualization* (2005).
3. MediaWiki API. https://www.mediawiki.org/wiki/API:Main_page. May 2016
4. Milgram, S. The Small World Problem. In *Psychology Today* (1967).
5. Reingold, E.M. and Tilford, J.S. Tidier Drawings of Trees. In *IEEE Transactions on Software Engineering.* (March 1981).
6. Wikipedia: Notability. https://en.wikipedia.org/wiki/Wikipedia:Notability