Applying Weighted PageRank to Friendships on Social Media Networks

CS 6850: The Structure of Information Networks

Github Repository

1. Introduction

In the past few decades, there has been growing interest in understanding social relationships within the lens of online social networks (Garton et al., 1997; Steinfield et al., 2013; Jurgens, D., 2013; Heer & Boyd, 2005; Arnaboldi et al., 2012; Gjoka et al., 2011). However, much of the research on analyzing social media networks has focused on binary relationships (Boyd & Ellison, 2007; Adamic & Adar, 2003), where users are either friends or not, with no consideration for the strength or nature of their connections. This binary concept of a "friend" on social media tends to oversimplify and abstract away crucial information about the true nature and degree of a friendship, which can limit our ability to understand and analyze social networks. A more meaningful metric for understanding friendships is relationship strength. In this context, relationship strength, or tie-strength, refers to the degree or intensity of a connection between two individuals on a social network (Granovetter, 1973).

This concept of relationship strength brings us to the primary research question of our project: How can we accurately and effectively model relationship strength between friends in a social network?

1.1 Literature Review

There are many approaches for modeling relationship strength in online social media platforms, but these existing models are often context-dependent and derived from available data, making it difficult to generalize across different social networks.

One notable paper that highlights this gap in the research is "Predicting Tie Strength with Social Media'' by Eric Gilbert and Karrie Karahalios, which presents a standardized method for measuring relationship strength between friends on social media platforms. In this paper, Gilbert and Karahalios acknowledge the shortcomings of previous tie strength research, which mainly centers around individual factors like communication frequency or emotional closeness, without taking into account broader social network structures or interaction patterns (Gilbert & Karahalios, 2009).

To address these limitations, Gilbert and Karahalios propose and develop a new model that uses social media data to predict tie strength between friends on social networks. This model is guided by 7 major dimensions of tie strength: Intensity, Intimacy, Duration, Reciprocal Services (Friedkin, 1980), Structural, Emotional Support (Wellman & Wortley, 1990), and Social Distance (Granovetter, 1973). This model for tie strength is a linear combination of selected predictor variables, each of which maps to one of the 7

tie strength dimensions, and includes terms that represent dimension interactions and network structure. The set of predictor variables selected for this model was purposely chosen for their common availability in most social media platforms so that they could generally be applied to any social network.

In order to evaluate the performance of this model, 35 participants were asked to rate the strength of their social media friendships by answering, “How strong is your relationship with this person?” on a scale of “barely know them” to “we are very close.” The results showed that this model fits the data very well. On average, this model predicted tie strength within one-tenth of its true value for the “How strong?” question. The predictive power of the 7 tie strength dimensions and their top three contributing variables was also analyzed. In this analysis, the Intimacy dimension made up the greatest contribution to tie strength (32.8%) followed by Intensity (19.7%).

This paper made two significant contributions to the existing literature:

1. It demonstrated the importance of **dimensions** of tie strength as manifested in social media which differed from previous studies that also quantified tie strength. Because each of the 7 dimensions contributed to the predictive power of the model, Gilbert and Karahalios conclude **that an accurate model should incorporate predictor variables from all 7 dimensions of tie strength.**
2. It differentiated from prior studies by proposing a model that is **platform-independent** and **generalizable** across different social media platforms. This is because the model was developed with a top-down approach. The authors first determined 7 core dimensions of tie strength, which were derived from the social science theory of friendships. From these 7 dimensions, the authors then selected specific feature variables from the data available that mapped to and spanned across these dimensions. **This process can be applied to any social media platform.**

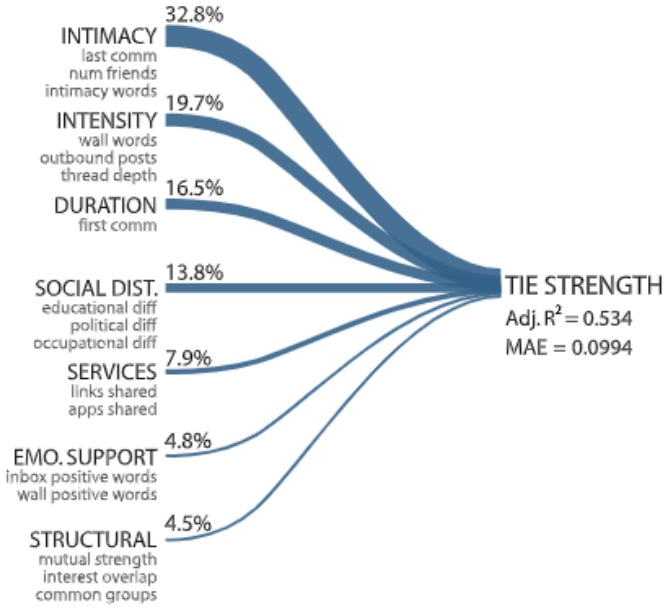

Figure 1: Illustration of the seven tie strength dimensions, their respective predictive power (Gilbert & Karahalios, 2009)

The results of the Gilbert and Karahalios paper open the door for large-scale social network analysis. As mentioned earlier in this section, there has been much research that measures tie strength in some way, but not to the extent of the model in this paper. We want to ground this idea in a concrete example, such as in the paper "The Role of Social Networks in Information Diffusion" by Eytan Bakshy. This paper concludes that though strong ties are individually more influential, weak ties are responsible for the propagation of novel information (Bakshy et al., 2012). Bakshy's research suggests that weak ties play a significant role in online information diffusion. But how does Bakshy actually model the strength of ties in his research? We observed that tie strength was measured in Bakshy's research using 4 types of predictor variables, all of which map only to the Intensity dimension of the Gibert Karahalios tie strength model. It is interesting to consider how leveraging the Gilbert Karahalios tie strength model, a more thorough measure, might provide new insight into this area of research, especially since the Intimacy dimension was found to have the most predictive power, significantly more than the Intensity dimension (Gilbert & Karahalios, 2009).

1.2 Project Goal

While the Gilbert and Karahalios model has proved to be effective in predicting tie strength between friends on social media, their study involved collecting survey data from individuals to determine the actual strength of friendships. Our project builds upon this framework with a more mathematical approach

The goal of our project is to investigate how modeling relationship strength using different heuristics will impact the results of network analysis algorithms, specifically the PageRank algorithm, when applied to friendship networks on social media.

To accomplish this goal, we:

1. Selected a social media network dataset 'YouTube' (Tang et al., 2009; Tang & Liu, 2009)
2. Developed a few different heuristics from this dataset to model tie strength
3. Ran the PageRank algorithm with modifications to include weights for each of these models

We can create different tie strength models by choosing heuristics from only one dimension of tie strength or choosing heuristics that span some combination of multiple dimensions. From here, we can examine whether the different models lead to different ranking outputs from PageRank and evaluate the impact of choosing different models. We can then observe how well the results from the algorithm align with the conclusions drawn by Gilbert and Karahalios.

Our methodology for accomplishing this goal is discussed in detail in the sections below.

2. Methodology

2.1 PageRank

PageRank is an algorithm used for ranking the importance of websites that was developed by Google's founders. PageRank served as the foundation for Google search's engine, and still plays a part in Google Search, though the modern-day ranking algorithm is certainly far more complex. The way PageRank determines the importance of a webpage is by counting the number and quality of links to that webpage. This is based on the idea that important webpages are more likely to be linked to by other webpages. To understand this better we can look at the mathematical formula of the original PageRank algorithm (Brin & Page, 1998; Rogers, 2002):

$$PR(A) = (1 - d) + d(\frac{PR(T1)}{L(T1)} + \frac{PR(T2)}{L(T2)} + \ldots + \frac{PR(TN)}{L(TN)}) \text{ , where}$$

- ***PR(A)*** is the PageRank (PR) of page A
- ***T1, T2, …., TN*** are the set of N pages that link to page A
- ***PR(T1)...PR(TN)*** are the PageRanks of each page T1, … TN that links to page A
- ***L*** is the number of outgoing links of a page
- ***d*** is the damping factor

We can see that the PageRank of webpage A depends on the PageRank of its inlinks, and the total number of webpages these inlinks point to. For example, if webpage T1 points to A and T2, half of its PageRank score will be transferred to A and half of its PageRank score will be transferred to T2. It is important to see here that the way "quality of links'' is assessed in PageRank is not through weighted edges, but through the importance of the webpages that link to your own webpage. PageRank only considers the topology of the graph, meaning the existence and direction of edges between web pages. There is nothing in the original PageRank to map to the concept of tie strength in friendships. For these reasons, we need a version of PageRank that takes into account weights. This led us to examine Weighted PageRank.

2.2 Weighted PageRank

The Weighted PageRank (WPR) algorithm, proposed by Wenpu Xing and Ali Ghorbani, is an extension of the traditional PageRank algorithm. The main difference between WPR and PR is that for a page A, Weighted PageRank assigns larger values to the more important links to page A instead of dividing the rank value of Page A evenly among all the pages that link to A. Each page that links to A gets a value proportional to its popularity, which is based on the number of its inlinks and outlinks. (Xing & Ghorbani, 2004). The formula for the Weighted Page Rank algorithm and a description of the variables represented in the formula is as follows (Xing & Ghorbani, 2004):

$$PR(A) = (1 - d) + d[\, PR(T1) * W_{in}(T1, A) * W_{out}(T1, A) + \ldots + PR(Tn) * W_{in}(Tn, A) * W_{out}(Tn, A)\,]$$

- ***PR(A)*** is the PageRank (PR) of page A

- ***T1, T2, …., TN*** are the set of N pages that link to page A
- ***PR(T1)...PR(TN)*** are the PageRanks of each page T1, … TN that links to page A
- ***d*** is the damping factor
- **W_{in}(T, A)**: see below
- **W_{out}(T, A)**: see below

W_{in}(T, A), in general, is the weight of the link from T to page A based on the number of inlinks of page A and the number of inlinks of all reference pages of page T

$$W_{in}(T, A) = \frac{I_A}{\sum_{p \in R(T)} I_p}$$

- I_A is the number of inlinks page A
- I_p is the number of inlinks page p
- R(T) is all the pages that v references (aka pages that T points to)

W_{out}(T, A), in general, is the weight of the link from T to page A based on the number of outlinks of page A and the number of outlinks of all reference pages of page T.

$$W_{out}(T, A) = \frac{O_A}{\sum_{p \in R(T)} O_p}$$

- O_A is the number of outlinks of page A
- O_p is the number of outlinks of page p
- R(T) is all the pages that T references (aka pages that v points to)

2.3 PageRank, with Weights

This exploration of PageRank and Weighted PageRank was necessary to assess how to achieve our main goal, which is to model relationship strength in a friendship network and analyze it using some network analysis algorithm. After reviewing the mathematics behind the Weighted PageRank algorithm, we observed that the algorithm does not divide the rank value evenly based on the number of outgoing links as in PageRank, and instead replaces this calculation with weights based on inlinks and outlinks. For our purposes, relationship strength (aka weights) should not depend on the network topology, like inlinks and outlinks. Furthermore, we want rankings to consider the number of outgoing links in conjunction with relationship strength, not to replace the former completely with the latter. Thus, we concluded that the Weighted PageRank algorithm is not the best approach for our purposes. Instead, it would be better to make a modification to the PageRank algorithm, which is PageRank but with weights. Though PageRank with weights does closely resemble Weighted PageRank, the key difference between the two is that in PageRank, weights are assigned on a per edge basis, while in Weighted PageRank, for each page, weights are assigned on a per outlink basis. The former more closely resembles the way we want to represent the relationship strength of a friendship as a weight assigned to an edge between two nodes.

To summarize in more concrete terms:
We decided against Weighted PageRank:

$$PR(u) = (1 - d) + d \sum_{v \epsilon B(u)} [PR(v) * W_{in}(v, u) * W_{out}(v, u)]$$

Instead, we chose to use PageRank:

$$PR(u) = (1 - d) + d \sum_{v \epsilon B(u)} [PR(v) \frac{PR(v)}{L_v}]$$

But **added a slight modification to PageRank to incorporate the notion of weights**, as follows

$$PR(u) = (1 - d) + d \sum_{v \epsilon B(u)} [PR(v) \frac{PR(v)}{L_v} * W(v, u)],$$

where W(v,u) is the weight of the link from v to u.

2.4 Applying PageRank with Weights to Friendship Relations, Friendship PageRank

In the section above, we compared different versions of PageRank, and decided that the algorithm most fitting for our project was PageRank, with a modification that includes weights. Although the PageRank algorithm is intended for measuring the importance of different websites on the Internet, our goal is to use it specifically for ranking friends with tie strengths. We can model PageRank with Weights in a friendship network with the following approach. From here on, we will refer to this mapping of PageRank with Weights to friendships on social networks as **Friendship PageRank.**

Just like in PageRank, Friendship PageRank's input is a directed graph. Nodes are people instead of web pages. In the PageRank algorithm, when web page i references web page j, an edge is added from node i to node j. Similarly, in Friendship PageRank, if A considers B to be a friend, an edge is added from node A to node B.

In order to incorporate the concept of tie strength, we need to modify PageRank with weights. In the original PageRank algorithm, the edges of the directed graph are unweighted. In Friendship PageRank, the edges of the graph are weighted, where the weight of each edge connecting two nodes in Friendship PageRank signifies the strength of the relationship.

The last part of PageRank we map to Friendship PageRank is the damping factor. In the original PageRank, the damping factor accounts for the behavior of a websurfer who does not always follow the outgoing links of a webpage. Sometimes a websurfer might get bored, leave the page, and arbitrarily pick a new page (Brin & Page, 1998; Rogers, 2002; Xing & Ghorbani, 2004). In Friendship PageRank, the damping factor represents the probability that a person will meet someone randomly in a completely new setting instead of through existing friendship relations. This helps to account for the fact that people are not limited to making friends through their existing connections, but can also make friends who have no overlap with their existing social circles.

Like PageRank, the output of Friendship PageRank is a dictionary that lists each node of the graph along with a normalized score. While the score of a node in PageRank represents the importance or relevance of a web page, the score for our Friendship PageRank represents how sociable a person is and is based on the strength of their relationships with other nodes in the graph. The node with the highest ranked score corresponds to the person with the strongest social network, having numerous strong relationships attested to by other nodes.

Lastly, in PageRank, a page's score is influenced by the scores of the pages that link to it (inlinks), but not by the pages it links to (outlinks). Therefore, if page A has a link to page B, but page B does not have a link to page A, only page B's score will be affected by this link. The implication that the direction of the link matters in PageRank also carries over to friendships. In Friendship PageRank, the edge from A to B can have a different weight than the edge from B to A, meaning that the relationship between two nodes **A** and **B** can be asymmetric. This is important, as in real life, the way that one person perceives a relationship may not always be reciprocated in the same way. Person A may think they are very close friends with Person B, but Person B may feel they are just acquaintances. This relationship is imbalanced, and a directed graph accounts for this by considering both directions of the links between nodes.

Now that we have defined the mapping from PageRank to Friendship PageRank with weights, the next step is to create different methods for calculating the weights of edges between friends on social media. This means devising different relationship strength models. Because we are looking at friends on social media platforms, it is implied that a friendship between A and B will always imply a friendship between B to A. The friendship is mutual, but the degree to which each person perceives the friendship does not have to be. However, for the specific models of relationship strength discussed below, the friendships are assumed to be mutual. In other words, in our models of relationship strength, the weight for edge A to B is always equal to the weight for edge B to A.

While previously we implemented PageRank with weights ourselves, we were able to leverage the built-in PageRank function in the NetworkX package, which includes an optional parameter for edge weights.

3. Dataset

In search of finding a suitable dataset for comprehensive analysis of our idea, we explored several publicly available datasets. The main issue with these datasets was that many of them lacked information about friendship connections and features of the friendship, making it difficult to accurately construct proper social network graphs of friendships and reasonably design relationship strength models for the friendships. (Leskovec & Mcauley, 2012; Yang & Leskovec, 2012).

We finally chose to use the 'Youtube' dataset created by Lei Tang and Huan Liu because it contained information about friendship connections and multiple types of interactions between users and friends. (Tang et al., 2009; Tang & Liu, 2009).

The ‘YouTube’ dataset contains information from 15,088 active user profiles, and captures five types of “interactions”. These five “interactions” are:

1. The contact network between the 15,088 users. The contact network indicates which of the 15,088 users are actually friends on Youtube.
2. The number of shared friends between two users.
3. The number of shared subscriptions between two users.
4. The number of shared subscribers between two users.
5. The number of shared favorite videos.

We first process this dataset by looking at interaction 1, the contact network. We use this contact network to construct edges for all the friendships in the Youtube dataset. Next, we use interactions 2-5 to develop heuristics for modeling relationship strength of friendships, and we only calculate relationship strength for the friendship edges derived from the contact network. The heuristics we chose are discussed in detail in the next section.

4. Heuristics for Modeling Tie Strength

Each of the heuristics below describes a different approach for calculating relationship strength. The relationship strength calculation for each friendship corresponds to the weight of an edge between two friends in the Youtube dataset.

Heuristic #1: Social Distance

For heuristic 1, relationship strength is modeled by calculating the normalized sum of interactions 3, 4, and 5, namely shared subscriptions, shared subscribers, and shared favorite videos. We chose to combine these three interactions into one heuristic because they are all variables that reflect the shared interests between two Youtube friends. All three of these variables map to the Social Distance dimension of the Gilbert and Karahalios tie-strength model. The social distance dimension is meant to embody factors such as socioeconomic status, education level, political affiliation, and race and gender. This is true because the types of videos a user likes, watches, and subscribes to often provides direct insight into the user’s demographics, so having these shared interests indicates shared social distance.

Heuristic #2: Structural

For heuristic 2, relationship strength between two friends on Youtube is modeled by calculating the number of shared friends, which is interaction 2. This directly maps to the Structural dimension of the Gilbert and Karahalios model, where one of the structural predictor variables is the number of mutual friends (Gilbert & Karahalios, 2009).

Heuristic #3: Linear Combination 50/50

For heuristic 3, relationship strength between two friends on Youtube is modeled by calculating a linear combination of heuristic 1 and heuristic 2. This means that half the weight is determined by the shared subscriptions, subscribers, and likes and half the weight is determined by shared friends.

Heuristic #4: Linear Combination, Based on Predictive Power

For heuristic 3, relationship strength between two friends on Youtube is modeled by calculating a linear combination of heuristic 1 and heuristic 2. However, instead of distributing the weight as 50/50 like in Heuristic #3, we distribute the weight based on the predictive power of the tie strength dimension that each heuristic represents, as shown in Figure 1. This means that weight is 0.138 x Heuristic 1 + 0.045 x Heuristic 2. This is because the Social Distance dimension has 13.8% of the predictive power, and the Structural dimension has 4.5% of the predictive power.

5. Results

Each of these heuristics are used to calculate weights representing relationship strength. We passed in the edge weights calculated by each of these heuristics into Friendship PageRank, ran the Friendship PageRank algorithm, and examined the ranking results of each heuristic.

To visualize and understand the differences between our four heuristics, we plotted several figures based on the ranking results.

5.1 Comparing Heuristic 1 to Heuristic 2

We first looked at the differences between Heuristic 1 and Heuristic 2. We anticipated that the differences would be large because the Social Distance dimension is completely separate from the Structural dimension. Figure 3 displays the variation in the node ranking position when switching from Heuristic 1 to Heuristic 2. Though it appears like there is a significant difference between the results of the two heuristics in this graph, the large amount of points we are plotting makes it difficult to interpret.

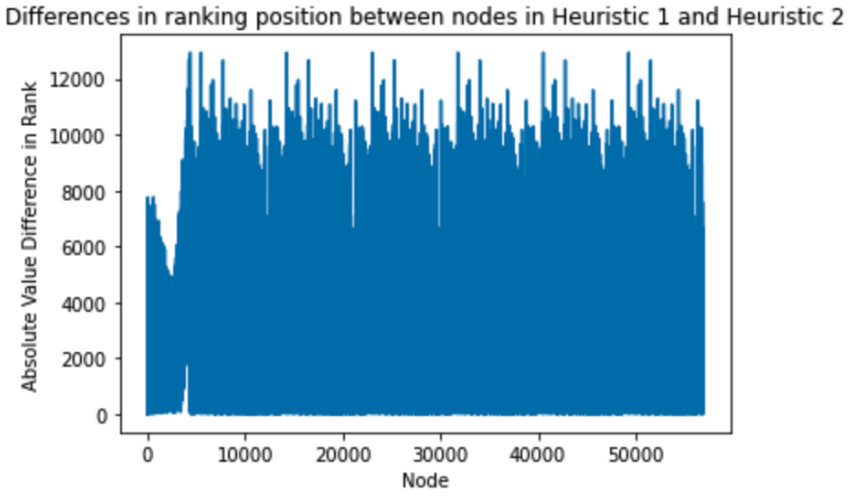

Figure 3: Distribution across all nodes of ranking differences when using Heuristic 1 versus Heuristic 2. The y-axis is the absolute value of the rank position of a node in Heuristic 1 minus the rank position of that corresponding node in Heuristic 2.

To address this, we looked at the distribution of the actual number of nodes with each difference, as shown in Figure 4. This figure shows that the many nodes did show a relatively small change in ranking between Heuristic 1 and Heuristic 2 ranking. However, the majority of nodes still had a significant difference between their Heuristic 1 and Heuristic 2 ranking. More specifically, for the majority of the nodes, their rankings are changed by thousands of positions when using Heuristic 1 compared to Heuristic 2.

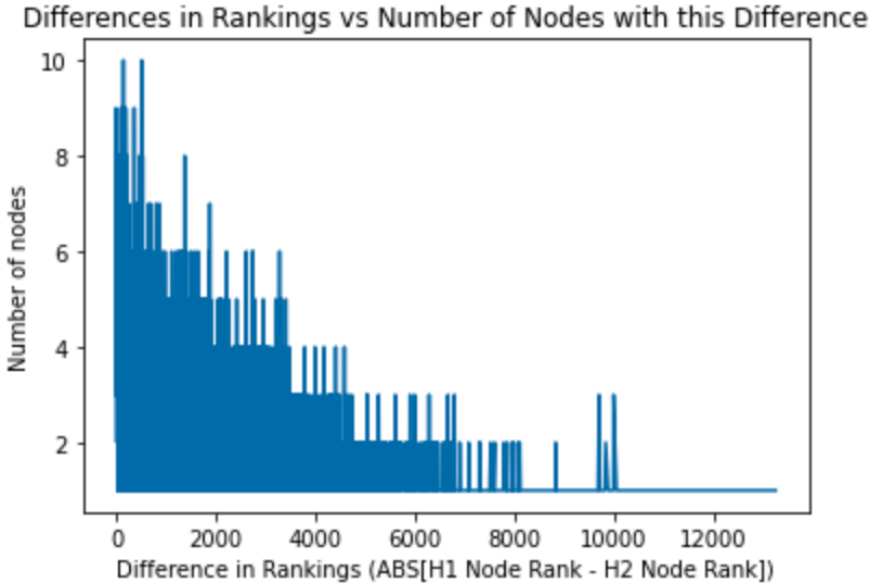

Figure 4: Differences in Ranking when using Heuristic 1 vs Heuristic 2.The y-axis is the difference in rankings while the x-axis is the respective number of nodes with such difference.

The results of this comparison between Heuristic 1 and Heuristic 2 show that the two models produce significantly different results.

5.2 Comparing Heuristic 1 to Heuristic 3

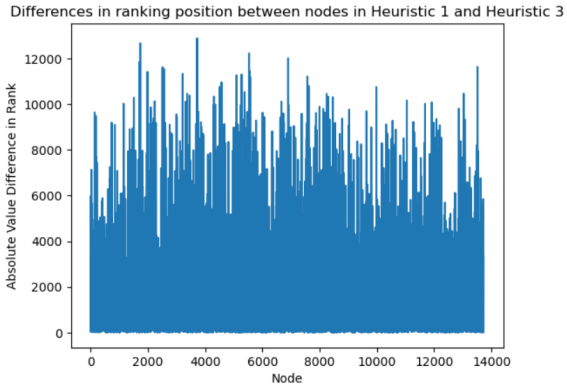

Figure 5: Distribution across all nodes of ranking differences when using Heuristic 1 versus Heuristic 3. The y-axis is the absolute value of the rank position of a node in Heuristic 1 minus the rank position of that corresponding node in Heuristic 3.

When comparing heuristic 1 and heuristic 3, similar to the comparison between heuristic 1 and heuristic 2 (as shown in Figure 5), we observe a significant difference in rankings for thousands of nodes. These differences range from 0 to 14,000 nodes, with some nodes experiencing a substantial rank difference of at least 4,000. Notably, a striking example is seen with nearly 14,000 nodes exhibiting an approximate rank difference of 11,000 between the two heuristics. Such a substantial change in rank highlights the magnitude of the impact caused by the variation in these heuristics.

5.3 Comparing Heuristic 2 to Heuristic 3

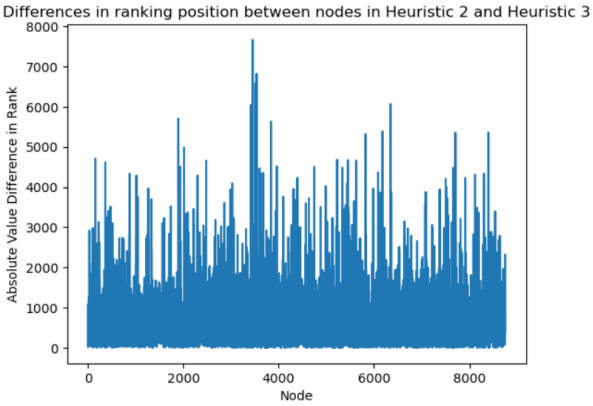

Figure 6: Differences in Ranking position when using Heuristic 2 vs Heuristic 3.The y-axis is the absolute value difference in rankings while the x-axis is the respective number of nodes with such difference.

When comparing heuristic 2 and heuristic 3, we observe a notable disparity in rankings across thousands of nodes, similar to the patterns we have previously discussed. However, the magnitude of this difference is not as significant as the disparity observed between heuristic 1 and heuristic 3. Figure 6 depicts this comparison, where approximately 8,000 nodes exhibit a rank difference of nearly 2,000. In contrast, Figure 5 showcases a more pronounced contrast, with approximately 14,000 nodes displaying a rank difference of around 11,000. Consequently, we can infer that while the results between heuristic 2 and heuristic 3 are not as markedly distinct as those between heuristic 1 and heuristic 3, they still demonstrate discernible variations.

5.4 Comparing Heuristic 3 to Heuristic 4

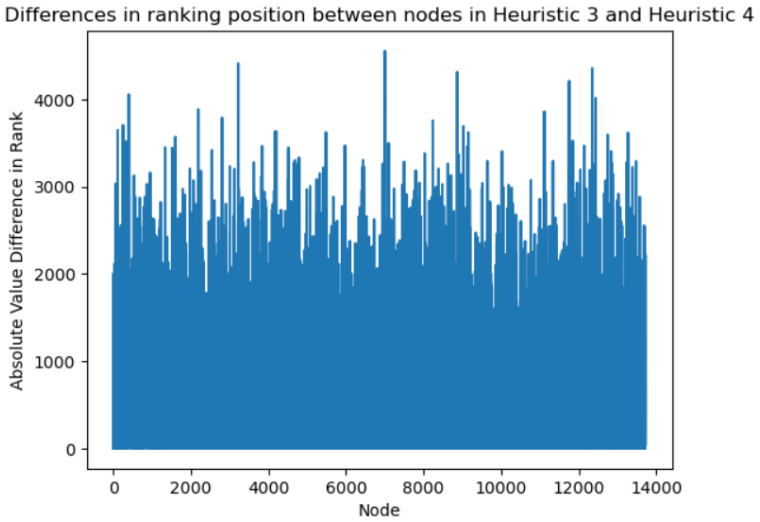

Figure 9: Differences in Ranking position when using Heuristic 3 vs Heuristic 4.The y-axis is the absolute value difference in rankings while the x-axis is the respective number of nodes with such difference.

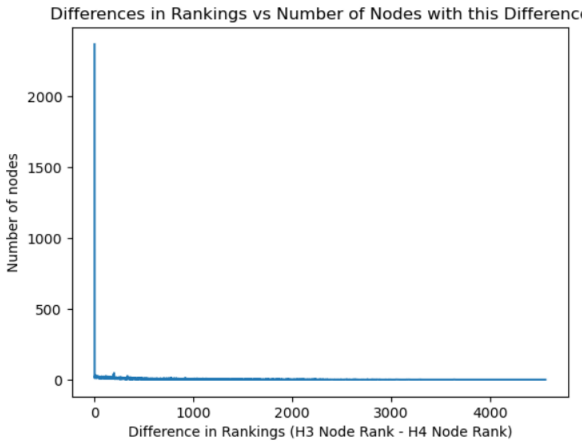

Figure 10: Differences in Ranking position when using Heuristic 3 vs Heuristic 4.The x-axis is the difference in rankings between both heuristics and the x axis is the accompanying number of nodes for such differences.

When comparing heuristic 3 and heuristic 4, we observe that the majority of nodes exhibit some variation in their ranking differences, consistent with the findings from other heuristic comparisons and as depicted in Figure 9. However, the magnitude of this change is relatively insignificant, as evidenced by Figure 10. In Figure 10, the largest bar graph represents the lowest difference in rankings, indicating that the majority of nodes experience minimal variation between heuristic 3 and heuristic 4. This trend is further supported by the overall decrease in the number of nodes on the y-axis as we progress along the x-axis, reflecting increasing differences in rankings. Consequently, it can be concluded that the majority of nodes demonstrate negligible changes in ranking differences between heuristic 3 and heuristic 4.

In conclusion, our analysis of different heuristics reveals varying degrees of changes in rankings. While some of these changes are significant, others are relatively minimal. Collectively, these findings support our overall conclusion.

Additional graphs comparing the different pairs of heuristics can be found in our Github Repository.

6. Conclusion

To summarize, the core question we tried to answer in this project was: Do the Friendship PageRank ranking results vary significantly across different relationship strength models? This would indicate that the specific choice of heuristic is crucial for calculating weights. Or, do the ranking results remain fairly similar across different models? This would indicate that any reasonable choice of heuristic to represent relationship strength could be appropriate when calculating weights.

The results comparing our different heuristics underscores the implication that heuristic choice is a crucial consideration when performing large scale analysis on a social network. Based on the results shown in Figures 3-8, we have observed that choosing a tie strength model with heuristics from only one dimension over another dimension, or from multiple dimensions instead of one dimension, can significantly change the outcomes of network analysis algorithms like PageRank. The results of our study prove to be a cautionary tale to the research world at large. When working on studies involving tie strength measurement, researchers should exercise careful consideration of their heuristic choices.

In addition, the results of our project align with the work of Gilbert and Karahalios, who emphasized the need for a standardized and accurate approach to modeling tie strength. The Gilbert and Karahalios tie strength model thoughtfully incorporates predictor variables across multiple dimensions of tie strength, and our research demonstrates the importance of following in their footsteps.

Future research in this field should continue to refine tie strength measurement models, striving for a standardized approach that captures the nuances of social relationships. However, this also opens up a different set of research questions for tie strength theory. To what extent can we actually model tie strength through data? Are there some aspects of human relationships that are inevitably lost when relying on predictive variables? What is the upper limit of tie strength predictability?

REFERENCES

Garton, L., Haythornthwaite, C., & Wellman, B. (1997). Studying online social networks. *Journal of computer-mediated communication*, 3(1), JCMC313.

Steinfield, C., Ellison, N. B., Lampe, C., & Vitak, J. (2013). Online social network sites and the concept of social capital. *Frontiers in new media research*, 122-138.

Jurgens, D. (2013). That's what friends are for: Inferring location in online social media platforms based on social relationships. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 7, No. 1, pp. 273-282).

Heer, J., & Boyd, D. (2005, October). Vizster: Visualizing online social networks. In *IEEE Symposium on Information Visualization*, 2005. INFOVIS 2005. (pp. 32-39). IEEE.

Arnaboldi, V., Conti, M., Passarella, A., & Pezzoni, F. (2012, September). Analysis of ego network structure in online social networks. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Conference on Social Computing* (pp. 31-40). IEEE.

Gjoka, M., Butts, C. T., Kurant, M., & Markopoulou, A. (2011). Multigraph sampling of online social networks. *IEEE Journal on Selected Areas in Communications*, 29(9), 1893-1905.

Boyd, D. M., & Ellison, N. B. (2007). Social network sites: Definition, history, and scholarship. *Journal of computer-mediated Communication*, 13(1), 210-230.

Adamic, L. A., & Adar, E. (2003). Friends and neighbors on the web. *Social networks*, 25(3), 211-230.

Granovetter, M. S. (1973). The strength of weak ties. *American journal of sociology*, 78(6), 1360-1380.

Gilbert, E., & Karahalios, K. (2009, April). Predicting tie strength with social media. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 211-220).

Friedkin, N. (1980). A test of structural features of Granovetter's strength of weak ties theory. *Social networks*, 2(4), 411-422.

Wellman, B., & Wortley, S. (1990). Different strokes from different folks: Community ties and social support. *American journal of Sociology*, 96(3), 558-588.

Bakshy, E., Rosenn, I., Marlow, C., & Adamic, L. (2012, April). The role of social networks in information diffusion. In *Proceedings of the 21st international conference on World Wide Web* (pp. 519-528).

Tang, L., Wang, X., & Liu, H. (2009, December). Uncovering groups via heterogeneous interaction analysis. In *2009 Ninth IEEE International Conference on Data Mining* (pp. 503-512). IEEE.

Tang, L., & Liu, H. (2009, May). Uncovering cross-dimension group structures in multi-dimensional networks. In *SDM workshop on Analysis of Dynamic Networks* (pp. 568-575).

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 30(1-7), 107-117.

Rogers, E. (2002). Diffusion of preventive innovations, *Addictive Behaviours*. Volume 27, 847-1048.

Xing, W., & Ghorbani, A. (2004, May). Weighted pagerank algorithm. In *Proceedings. Second Annual Conference on Communication Networks and Services Research*, 2004. (pp. 305-314). IEEE.

Leskovec, J., & Mcauley, J. (2012). Learning to discover social circles in ego networks. *Advances in neural information processing systems*, 25.

Yang, J., & Leskovec, J. (2012, August). Defining and evaluating network communities based on ground-truth. In *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics* (pp. 1-8).