Predicting Urban Flow Through PageRank: A Comparative Evaluation

Lilian Huang, Katy Koenig, Nora Hajjar

1 Introduction

This project investigates the PageRank algorithm through applying it to an unconventional domain area: predicting the flow of urban traffic. Our objective is twofold. First, by using the PageRank algorithm to analyze Transportation Network Provider ("rideshare") trips in Chicago through treating community areas as nodes, and the trips between them as edges, we compute and evaluate an estimate of which community areas are most-frequented and can therefore be regarded as most important in the network of Chicago rideshare commuter patterns. Besides exploring this specific dataset, we also investigate the properties of the PageRank algorithm itself: the distinction between regular and weighted PageRank, and the effect of the damping parameter upon the accuracy of the final estimate.

1.1 Literature review

Gleich's 2015 paper, *PageRank Beyond the Web*, examines how PageRank has been used in various domains to determine the relative importance of each member in a network of relationships. Gleich highlights that one of the more unusual applications of the algorithm is to predict traffic flow in road networks, as done by Jiang, Zhao, and Yin (2008), who found that PageRank was one of the best-performing metrics for doing so.

Subsequent studies (Jiang, 2009; Jiang, Yin and Zhao, 2009) further explored using different variants of the PageRank algorithm to model these traffic flows, including comparing the performance of the regular PageRank algorithm with that of the weighted PageRank algorithm. In regular PageRank, once the random walker is at a certain node i, the probability of transitioning from node i to any of its neighbors (linked nodes) is equal for each neighbor j; in contrast, the random walker in weighted PageRank has a preference for transitioning to high-degree neighbors (neighbors which have a large number of in-links and out-links, relative to the total number of links for all of i's neighbors) - in other words, there is a prioritization of certain nodes based on their popularity relative to their counterparts.

The regular PageRank formula is:

$$PR(i) = \frac{1-d}{N} + d \sum_{j \in M(i)} \frac{PR(j)}{n_j}$$

where $PR(i)$ and $PR(j)$ are the PageRank scores of nodes i and j, N is the total number of nodes, $M(i)$ is the set of nodes which link to node i, n_j is the number of outbound links from node j (i.e. nodes to which node j links), and d is a damping parameter.

The weighted PageRank formula (Xing and Ghorbani, 2004) is:

$$WPR(i) = \frac{1-d}{N} + d \sum_{j \in M(i)} WPR(j) \frac{w(i)}{\sum w(k)}$$

The final term $\frac{w(i)}{\sum w(k)}$ is the weight applied, which can be defined as node i's relative popularity among its counterpart nodes (the other nodes to which node j links); $w(i)$ is the total number of in-links and out-links for node i, and $\sum w(k)$ is the total number of in-links and out-links for all nodes to which node j links.

Both the regular and weighted PageRank algorithms contain a damping parameter d that represents the probability of continuing to follow links from the current node; this is typically set to 0.85, by Google's precedent. $1 - d$ represents the probability of the random walker "teleporting" to a random node.

Jiang's analysis briefly notes that weighted PageRank was found to outperform regular PageRank in modeling traffic flows.

Our project builds upon this body of existing research in several ways: firstly, by applying the PageRank algorithm to a new and recent dataset; the preexisting studies have focused on predicting pedestrian and vehicle traffic along specific road segments, whereas we seek to understand the flow between neighborhoods. Secondly, we are furthering the exploration of how regular PageRank compares to weighted PageRank in terms of prediction quality, as it applies to urban networks, to verify if Jiang's suggestion holds.

In addition, Jiang (2009) observes that for higher values of parameter d, PageRank scores are more strongly correlated with actual traffic flow, but offers minimal elaboration or reasoning as to why, and does not specify what constitutes a high value of d. We investigate this by systematically running the algorithm with varying values of d and examining the effect on calculated PageRank scores.

2 Methodology

2.1 Data

We use a dataset of trip reports by Transportation Network Providers, i.e. rideshare companies, publicly available from the City of Chicago, which contains information on the community areas where each trip began and ended. Chicago contains 77 community areas, which are well-defined clusters of neighborhoods used for official policy and planning purposes.

The data is available from November 2018 through October 2019, but for feasibility, we limit our analysis temporally, using the single month of May 2019 as "training" data (for computing our PageRank scores), and the week of June 1, 2019 to June 7, 2019 as "testing" data (against which to compare our PageRank scores). Figure 1 shows the number of trips that pick up from and drop off at each community area within this time period.

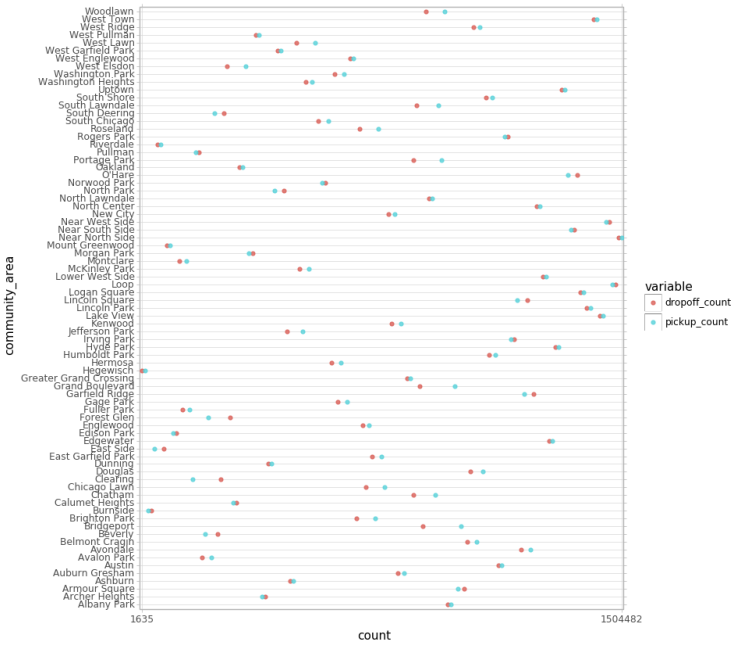

Figure 1: Descriptive plot of trip pickups and dropoffs.

2.2 Evaluation metrics

We compute PageRank scores using both regular and weighted PageRank algorithms, applied to a graph of rideshare trips between all 77 Chicago community areas. Since the vector of PageRank scores represents the probability that a random walker will be at a particular node (community area), once the algorithm converges, we evaluate these scores by comparing them to the actual proportion of trips which end in each community area, in the testing data.

In line with previous literature evaluating PageRank performance (Chin and Wen, 2015; Son, Christensen, Grassberger, and Paczuski, 2012), we carry out this comparison by calculating the correlation coefficient between PageRank scores and the reference values. We use the Kendall's tau correlation coefficient, since we are working with non-normal data, and it allows us to evaluate relative rank - PageRank is an ordinal property, and as such we are more interested in the ordering of the scores, rather than their absolute values.

3 Results and Discussion

3.1 Comparison of regular and weighted PageRank

We make an initial comparison of the performance of the regular PageRank and weighted PageRank algorithms. We see that weighted PageRank displays a substantially higher correlation with the observed TNP traffic flow - the correlation coefficient is notably higher than the correlation coefficient for the regular PageRank scores and observed TNP traffic flow.

d value	0.15	0.50	0.85
Regular PageRank	0.593399	0.590524	0.592675
Weighted PageRank	0.654589	0.705862	0.827551

Figure 2: PageRank correlation with actual traffic flow, for regular versus weighted PageRank, at different values of the damping parameter.

This indicates that the weighted PageRank algorithm may be more suitable for predicting actual traffic flow in this dataset. This is most likely because weighted PageRank is more appropriate for simulating the setup of vehicle traffic, or indeed any form of commuter movement within physical space. Under regular PageRank, the probability of following a link to a different node is determined by simple connectivity (whether a link/trip exists between the two nodes or not), i.e. a random walker has equal probability of transitioning to any of the adjacent nodes. However, weighted PageRank takes into consideration how highly connected each adjacent node is overall, and this is a better reflection of the situation - in physical movement, there is a preference for highly-connected nodes, because there are more physical routes (e.g. streets) available to reach them, and potentially because of network effects; people may have more incentive to visit community areas that receive high traffic from other people. In other words, in this context, a highly-connected node (a community area with a large number of trips) is likely to offer greater convenience (more routes to arrive at it) and greater attractiveness (more reasons to visit it).

3.2 Effect of damping parameter *d*

We also examine the effect of varying the damping parameter d beyond the standard value of 0.85.

3.2.1 Effect on correlation with actual values

From Figure 3.1, we can see that the previous observation persists - weighted PageRank outperforms regular PageRank in terms of its correlation with actual data, for all values of the damping parameter. Furthermore, this gap grows more pronounced as the damping parameter increases; the correlation between weighted PageRank and actual traffic flow becomes increasingly strong as the damping parameter increases. While regular PageRank also improves in performance, this change is very slight and much less perceptible.

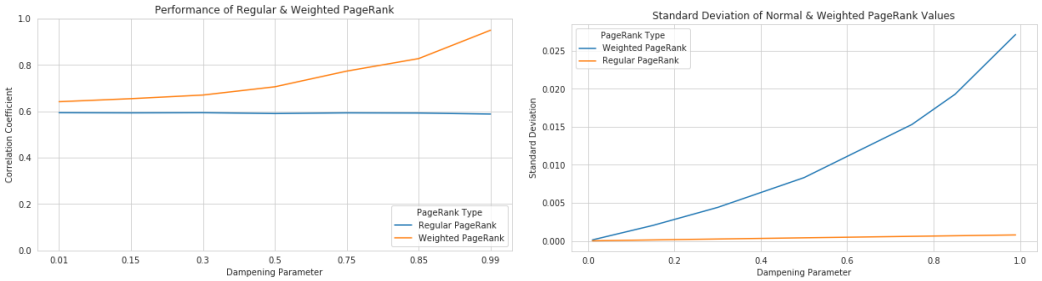

Figure 3: (Left, Figure 3.1) The correlation of regular and weighted PageRank with actual traffic flow, for different values of d. (Right, Figure 3.2) The standard deviation of the PageRank vector, for different values of d.

To understand why a larger damping parameter improves predictive performance, we must recall that $1 - d$ can be interpreted as the likelihood of a random walker "teleporting" rather than continuing to follow links from the node. In other words, the smaller the damping parameter d, the more the simulation is dominated by random "teleportation", which is not a good model for physical movement. A possible example of random "teleportation", in this model, might be equivalent to switching to a different (non-rideshare) form of transportation to continue one's journey, which is much more heavily constrained than an Internet user randomly "teleporting" to a different webpage. A larger damping parameter means that there is a greater probability of the random walker following a long chain of links between nodes, without randomly jumping, which better models this particular situation.

Previous research has noted that choosing a large damping parameter (a value close to 1) can in fact lead to poor predictive performance, as random walkers become concentrated in "ranksinks" - nodes which have incoming links but no outgoing links - resulting in PageRank scores that are not reflective of the nodes' actual importance; under these circumstances, many important nodes may have misleadingly low PageRank. However, our findings indicate that this may be more of a concern with the conventional Internet-based model of PageRank, rather than when it is applied to the domain of physical traffic flows. As such, when dealing with other (non-Internet) domain areas, it may be less crucial to strike a balance between having the damping parameter be too small or too large.

3.2.2 Effect on PageRank distribution itself

We also examine how changing the damping parameter affects the PageRank vector itself. We can see from Figure 3.2 that overall, there is minimal variation in the PageRank vector for regular PageRank, i.e. the PageRank scores for the community areas are all very similar. However, we see that for weighted PageRank, the standard deviation of the PageRank vector increases as the damping parameter increases; this is because, when the damping parameter is 0, only random "teleportation" occurs, and so every PageRank score is equal to 1/N (where N is the number of nodes).

4 Conclusion

One key result is that weighted PageRank performs better at modeling traffic flows than regular PageRank, as evaluated by how well the PageRank scores are correlated with actual traffic flows. Furthermore, the strength of this correlation increases with the value of the damping parameter.

Future research could further explore which metrics are most suitable for quantifying and evaluating the performance of the PageRank algorithm - for example, there are multiple possible correlation measures, and we only made use of one (the Kendall correlation coefficient). Another potential expansion would be more thoroughly investigating how variation in the damping parameter can cause both fluctuation in raw PageRank scores and "reversal" of relative rank, and the extent to which these changes occur in tandem. For example, instead of simply seeing how variation in damping parameter affects the PageRank vector's correlation with actual data, researchers can examine the correlation between two different PageRank vectors, computed using two different values of the damping parameter, to determine the extent to which the rankings have been "reversed" or destabilized. These are but a few of many promising avenues.

References

Gleich, D. F. (2015). *PageRank beyond the Web. SIAM Review, 57*(3), 321–363. doi:10.1137/140976649

Jiang, B. (2009). *Ranking spaces for predicting human movement in an urban environment. International Journal of Geographical Information Science, 23*(7), 823–837. doi:10.1080/13658810802022822

Jiang, B., Zhao, S., & Yin, J. (2008). *Self-organized natural roads for predicting traffic flow: a sensitivity study. Journal of Statistical Mechanics: Theory and Experiment, 2008*(07), P07008. doi:10.1088/1742-5468/2008/07/p07008

Jiang, B., Yin, J., & Zhao, S. (2009). *Characterizing the human mobility pattern in a large street network. Physical Review E, 80*(2), 021136.

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

Son, S. W., Christensen, C., Grassberger, P., & Paczuski, M. (2012). *PageRank and rank-reversal dependence on the damping factor. Physical Review E, 86*(6), 066104.

Chin, W. C. B., & Wen, T. H. (2015). *Geographically modified PageRank algorithms: identifying the spatial concentration of human movement in a geospatial network. PLoS ONE 10*(10): e0139509. https://doi.org/10.1371/journal.pone.0139509

Bressan, M., & Peserico, E. (2010). *Choose the damping, choose the ranking?. Journal of Discrete Algorithms, 8*(2), 199-213.