Visualization and Pagerank Analysis in Metropolitan Subway Network

Mulkyeol Kim
Department of Industrial Engineering
UNIST
ccomo456@unist.ac.kr

Jihwan Oh
Department of Industrial Engineering
UNIST
dhwlghks00@unist.ac.kr

Abstract

Our goal is to analyze Seoul and metropolitan subway pagerank. First, pre-processing of subway data in Seoul and the metropolitan area and Exploration data analysis are performed. Calculate the page rank using only a graph with no direction. But the results were almost random.

To solve this problem, the second task is to calculate the weighted pagerank using a direct graph. To calculate the weighted pagerank, the number of people getting on and off is estimated, and the weighted pagerank is calculated through the maximum eigenvalue method. The results were quite consistent with our common sense.

Additionally, correlation and OLS analysis were conducted, and future research directions were considered.

I. INTRODUCTION

A. Problem Definition

The metropolitan subway, which spreads from Seoul to Gangwon-do, Incheon, and Chungcheongbuk-do, etc., has a very large number of users so also very complicated. Due to the large number of passengers, they are very crowded, and safety and discomfort problems have been steadily raised. Considering these points, we thought that selecting highly important stations on the metropolitan subway would help us intensively take care of the station and solve those problems.

Therefore, our goal is to find stations of high importance. These stations play an important role between other stations and lines, and there should be a lot of passenger movements. We selected 'pagerank' as the standard. Thus, in summary, our project aims to analyze pagerank in metropolitan subways and visualize it for easy use(developing important subway stations become safer and more convenient) in the real world.

B. Hypothesis

Before starting the task, we hypothesized which inverse would be of high importance.

(1) The pagerank value of the station in central Seoul, not the outskirts, will be higher.
(2) The pagerank at the transit station will be generally high.

II. LITERATURE REVIEW

A. Subway Stations Network Structure Analysis by Using Social Network Analysis

In this literature, degree centrality, closure centrality, and eignvector centrality were calculated using the data on the number of people getting on and off the metropolitan subway. As a result of Literature's analysis, 'Gangnam' Station was ranked high not only in the number of people getting on and off but also in the centrality analysis, which can be seen as a hub in the subway station network. In addition, 'Jamsil' Station showed a high ranking in the number of getting on and off rankings, but the ranking fell in the centrality analysis. This indicates that 'Jamsil' Station has a large number of people getting on and off, but is connected to stations that are relatively insignificant.

Literature mentioned that even if one centrality index is high or low, other centrality indices are low or high, so it is necessary to analyze the characteristics of each station in more detail.' Therefore, when analyzing the subway network, it seems necessary to calculate the centrality or importance by considering not only one indicator, but also the relationship with the neighbor stations and the information on the number of people getting on and off.

B. Using Complex Numbers in Website Ranking Calculations: A Non-ad Hoc Alternative to Google's PageRank - Keita Sugihara

According to this paper, the power-iteration method is not available in complex networks. As an alternative to this, there is a method using weight matrix's eigenvector and eigenvalue. The page rank is also dependent on the maximum eigenvalue. Therefore, the page rank converges to the eigen vector when it has the maximum eigenvalue. Therefore, since our data is a complex network structure, we will use a method of finding the maximum value of the eigen value of the weight matrix instead of the power-iteration method that requires repetitive computation.

The PageRank formula as shown below, Equation 1.

$$\overrightarrow{\boldsymbol{pr}} = \vec{\boldsymbol{v}} \in \{ \boldsymbol{W}\vec{\boldsymbol{v}} = \boldsymbol{max}(\boldsymbol{\lambda})\vec{\boldsymbol{v}} \}$$
$$(\lambda : eigenvalue,\ \vec{v} : eigenvector\ of\ W) \quad (1)$$

III. METHOD

A. Data

We needed data with information about metropolitan subway lines (which should also include line order) and data with information on the number of people getting on and off. We were able to receive such data from ‘서울 열린 데이터 광장’, an open-source data site provided by the Seoul Metropolitan Government.

	전철역코드	전철역명	전철명명(영문)	호선	외부코드
0	0245	신답	Sindap	02호선	211-2
1	0336	학여울	Hangnyeoul	03호선	346
2	1014	청량리	Cheongnyangni	경의선	K117
3	1218	원덕	Wondeok	경의선	K136
4	1264	홍대입구	Hongik Univ.	경의선	K314
...	...	...	...	...	...
762	0159	동묘앞	Dongmyo	01호선	127
763	0200	까치산	Kkachisan	02호선	234-4
764	0201	시청	City Hall	02호선	201
765	0202	을지로입구	Euljiro 1(il)-ga	02호선	202
766	0300	대곡	Daegok	경의선	K322

767 rows × 5 columns

Fig. 1 Line data

	사용일자	노선명	역명	승차총승객수	하차총승객수	등록일자
20221001	3호선	고속터미널	59124	62989	20221004	NaN
20221001	3호선	교대(법원.검찰청)	8040	4875	20221004	NaN
20221001	3호선	학여울	3355	3401	20221004	NaN
20221001	3호선	대청	6517	5926	20221004	NaN
20221001	3호선	일원	6231	6025	20221004	NaN
...	...	...	...	...	...	...
20221031	6호선	버티고개	2366	2239	20221103	NaN
20221031	6호선	약수	3778	3751	20221103	NaN
20221031	6호선	청구	3842	4067	20221103	NaN
20221031	6호선	신당	8136	8814	20221103	NaN
20221031	6호선	동묘앞	8434	8751	20221103	NaN

18785 rows × 6 columns

Fig. 2 Data about number of people getting on and off

B. Data Preprocessing

(1) Selection of key lines

There are 24 subway lines in the metropolitan area, and some lines are centered in the outskirts (Gyeongchun Line), or have a special purpose(Eveline). So, we thought that using all 24 routes was inefficient and interfering with visualization and analysis, so we selected only lines 1-9, which people usually use for traffic purposes.

(2) Get information of pair between previous station and the next station is required

For the visualization of the graph of the metropolitan subway, we needed information on the pair of previous and next stations. The preprocessing was performed using the line data in Figure 1. Since column ‘외부코드’ contains information related to the line and its order, the station of the next row is designated in the ‘다음역’ column after sorting based on the ‘외부코드’. In addition, ‘다음역’ of the last station of each line was designated as the previous station so that they could be recognized as the end point of line.

(3) Branches coming out from a single line

Some one station had two next stations at the same time, forming a branch. So these was handled.

(4) Specify latitude and longitude using GoogleMap

To plot the station accurately, we used latitude and longitude as coordinates. For this, we used Googlemap. It finds latitude and longitude by name. By loading and utilizing Google Maps from Python, we could quickly and efficiently find latitude and longitude values for the names of numerous stations in data frame. We used it and created a data frame containing latitude and longitude.

	위도	경도
소요산역	37.947099	127.060681
동두천역	37.926664	127.054992
보산역	37.914277	127.057158
동두천중앙역	37.901673	127.056409
지행역	37.889979	127.064305
...	...	...
병점역	37.206821	127.033268
신도림역	37.50881	126.891206
성수역	37.544577	127.055991
강동역	37.53594	127.132187
응암역	37.598472	126.915592

456 rows × 2 columns

Fig. 3 Dataframe of latitude and longitude

Afterwards, when the data frame of Figure 1 and the data frame of Figure 3 are combined, a data frame containing comprehensive information such as the next station, latitude, longitude, and arc can be generated.

(5) Adding getting in and out data

df_people

	Date	호선	전철역명	승차총승객수	하차총승객수
0	20221001	3호선	고속터미널	59124	62989
1	20221001	3호선	교대(법원.검찰청)	8040	4875
2	20221001	3호선	학여울	3355	3401
3	20221001	3호선	대청	6517	5926
4	20221001	3호선	일원	6231	6025
...	...	...	...	...	...
18780	20221031	6호선	버티고개	2366	2239
18781	20221031	6호선	약수	3778	3751
18782	20221031	6호선	청구	3842	4067
18783	20221031	6호선	신당	8136	8814
18784	20221031	6호선	동묘앞	8434	8751

18785 rows × 5 columns

Fig. 4 Dataframe of number of people getting on and off

The data frame in Figure 4 is data that records the number of people getting on and off of each subway station by day. It is time series data recorded in a daily frequency.

In the data frame obtained in III.Methods-B-(4) above, value of '승차총승객수' and '하차총승객수' in dataframe of Figure2 is filled based on the name of the train station.

df_merged

	전철역명	전철명명(영문)	다음역	다음역(영문)	위도	경도	Date	승차총승객수	하차총승객수
0	소요산역	Soyosan	동두천	Dongducheon	37.947099	127.060681	20221001	5811	4981
1	소요산역	Soyosan	동두천	Dongducheon	37.947099	127.060681	20221002	4293	3936
2	소요산역	Soyosan	동두천	Dongducheon	37.947099	127.060681	20221003	2225	1708
3	소요산역	Soyosan	동두천	Dongducheon	37.947099	127.060681	20221004	2503	1871
4	소요산역	Soyosan	동두천	Dongducheon	37.947099	127.060681	20221005	2863	2178
...	...	...	...	...	...	...	...	...	...
18355	강동역	Gangdong	둔촌동	Dunchon-dong	37.53594	127.132187	20221027	20376	19609
18356	강동역	Gangdong	둔촌동	Dunchon-dong	37.53594	127.132187	20221028	20537	19406
18357	강동역	Gangdong	둔촌동	Dunchon-dong	37.53594	127.132187	20221029	15208	14552
18358	강동역	Gangdong	둔촌동	Dunchon-dong	37.53594	127.132187	20221030	10192	9856
18359	강동역	Gangdong	둔촌동	Dunchon-dong	37.53594	127.132187	20221031	19718	19120

18360 rows × 9 columns

Fig. 5 Final Dataframe

The data on the number of people getting on and off are completely combined.

There were also about five stations without getting on and off data, and according to our search, they were newly established in 2022 (within the last year). Since it was added to the 'end' of the subway line, they are in outskirts, and if the data on getting on and off is randomly filled or designated as a small number such as 1, they could affect the analysis of other stations, so these stations were removed.

C. Global Clustering Coefficient

Clustering coefficient refers to how connected some node's neighbors are to each other (relative to how connected they could be). If the clustering coefficient is calculated for the entire network, the global clustering coefficient can be obtained.

$$C = \frac{number\ of\ closed\ triplets}{number\ of\ all\ triplets} = \frac{3 \times \text{number of triangles}}{number\ of\ all\ triplets} \quad (2)$$

Global cluster coefficient can be calculated by Equation

D. NetworkX

NetworkX is a Python library for studying graphs and networks. NetworkX makes it easy to visualize graphs with large number of nodes and edges. In addition, the location (coordinate; 'pos') of nodes can be specified by the user, so it is possible to accurately plot the location of the stations using the latitude and longitude information obtained above. Also, it provides function which can calculate pagerank.

E. Pagerank

Pagerank, developed by Larry Page and Sergey Brin (founders of Google) is a method for rating the importance of web pages objectively and mechanically using the link structure of the web. Pagerank can be calculated by 'Power iteration'.

(1) Power iteration

Assume there is a graph with n nodes. First, an initial page rank is assigned each node. And it is repeated until convergence to equation (3).

$$\sum_i \left| r_i^{(t+1)} - r_i^{(t)} \right| < \varepsilon \quad (3)$$

Then when it converges, stop repeating and r_i is final pagerank.

(2) Weight estimation

<u>Case1: Normal station</u>

In normal station case, directed graph can be drawn as Fig6. Then, we focus on vector b_a and b_c that highlighted. We can estimate two values.
A_{in} and A_{out} means that number of passengers getting in and out at station A. We consider the weight using number of getting off passengers in both sides.

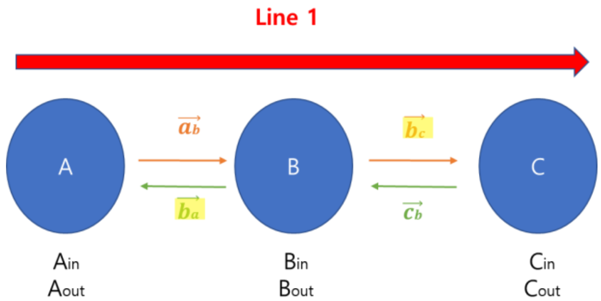


Fig6. The directed graph of normal station

We can estimate the passenger's movement. A_{in} and A_{out} means that number of passengers getting in and out at station A. The total amount of movement at station B is the same as B_{in}.

$$|\vec{b_c}| = \boldsymbol{B_{in}} \times \frac{\boldsymbol{C_{out}}}{\boldsymbol{A_{out}} + \boldsymbol{C_{out}}}$$

$$|\vec{b_a}| = \boldsymbol{B_{in}} \times \frac{\boldsymbol{A_{out}}}{\boldsymbol{A_{out}} + \boldsymbol{C_{out}}}$$

$$|\overrightarrow{\boldsymbol{b_{out}}}| = |\vec{b_c}| + |\vec{b_a}| = \boldsymbol{B_{in}} \tag{4}$$

Equation 4 is the estimated weight equation in normal station.

<u>Case2: Transfer station</u>

In transfer station case, directed graph can be drawn as Fig8. Then, we focus on vector b_a, b_c, b_d and b_e that highlighted. We can estimate four values.

We consider the weight using number of getting off passengers in both sides.
The total amount of movement at station B is the same as B_{in}.

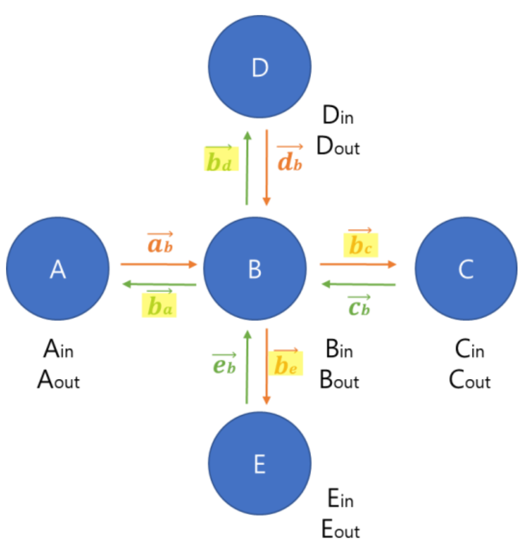


Fig7. The directed graph of normal station

We can estimate the passenger's movement as Equation 5.

$$|\vec{b_c}| = \boldsymbol{B_{in}} \times \frac{\boldsymbol{C_{out}}}{\boldsymbol{A_{out}} + \boldsymbol{C_{out}} + \boldsymbol{D_{out}} + \boldsymbol{E_{out}}}$$

$$|\vec{b_e}| = \boldsymbol{B_{in}} \times \frac{\boldsymbol{E_{out}}}{\boldsymbol{A_{out}} + \boldsymbol{C_{out}} + \boldsymbol{D_{out}} + \boldsymbol{E_{out}}}$$

$$|\vec{b_a}| = \boldsymbol{B_{in}} \times \frac{\boldsymbol{A_{out}}}{\boldsymbol{A_{out}} + \boldsymbol{C_{out}} + \boldsymbol{D_{out}} + \boldsymbol{E_{out}}}$$

$$|\vec{b_d}| = \boldsymbol{B_{in}} \times \frac{\boldsymbol{D_{out}}}{\boldsymbol{A_{out}} + \boldsymbol{C_{out}} + \boldsymbol{D_{out}} + \boldsymbol{E_{out}}}$$

$$|\overrightarrow{\boldsymbol{b_{out}}}| = |\vec{b_c}| + |\vec{b_e}| + |\vec{b_a}| + |\vec{b_d}| = \boldsymbol{B_{in}} \tag{5}$$

(3) Weight matrix
We can calculate all weight through weight estimation. Now, we can derive weight matrix W. (Fig.8)
Dimension of W is (387,387) because the number of all station is 387.

$$W =$$

```
[[0.         0.32088981 0.         ... 0.         0.         0.        ]
 [0.08804148 0.         0.         ... 0.         0.         0.        ]
 [0.         0.         0.         ... 0.         0.         0.        ]
 ...
 [0.         0.         0.         ... 0.         0.         0.        ]
 [0.         0.         0.         ... 0.         0.         0.        ]
 [0.         0.         0.         ... 0.         0.         0.        ]]
```

$$W \in R^{387X387}$$

Fig8. Weight matrix for all station.

(4) Weighted PageRank

We already know the power iteration method. But in our model, there are many edges and nodes. And the graph is very complicate. So, it does not converge. So, we can't derive the PageRank using power iteration. So, I introduce alternative method for find PageRank using Eigenvector and eigenvalues of $\boldsymbol{W}$. Another definition of PageRank is that eigenvector that maximum eigenvalue of $\boldsymbol{W}$. The result maximizes eigenvalue method is almost same as iteration method.

$$\overrightarrow{\boldsymbol{pr}} = \vec{\boldsymbol{v}} \in \{\, \boldsymbol{W}\vec{\boldsymbol{v}} = \boldsymbol{max(\lambda)}\vec{\boldsymbol{v}} \,\}$$

$$(\lambda : eigenvalue,\ \vec{v} : eigenvector\ of\ W) \tag{6}$$

Equation 6 is Weighted PageRank equation by maximizes eigenvalue.

Our PageRank result as Fig9.
The dimension of PageRank is (387,1).

$$\overrightarrow{pr} = \begin{bmatrix} 2.14066523e\text{-}07 \\ 2.87628408e\text{-}07 \\ 9.55609097e\text{-}08 \\ -3.55190359e\text{-}15 \\ 1.84967497e\text{-}07 \\ -2.98593169e\text{-}15 \\ \vdots \\ 2.31564763e\text{-}16 \\ 3.91016721e\text{-}16 \\ -1.82989102e\text{-}15 \\ 1.62645738e\text{-}08 \\ 3.00185305e\text{-}16 \\ 3.45165178e\text{-}16 \end{bmatrix} \in \boldsymbol{R^{387X1}}$$

Fig9. Weighted PageRank results

F. Correlation Matrix

Analyze the correlation between weighted PageRank and dependent variables. The dependent variables areLongitude, Latitude, Line, Total number of passengers getting on, Total number of passengers getting off.

G. Linear Regression (OLS)

Ordinary Least Squares regression (OLS) is a common technique for estimating coefficients of linear regression equations which describe the relationship between one or more independent quantitative variables and a dependent variable. Equation 7 shows that Weighted PageRank OLS equation.

$$\boldsymbol{pagerank} = \alpha(Longitude) + \beta(Latitude) + \gamma(Line) + \delta(\#\_of\_passengers) + \varepsilon$$

IV. Application

First, we can find the importance of subway stations through pagerank and visualizing them to identify at a glance. After this, it is possible to focus on expanding or improving the congested transfer procedure in stations with high importance. In fact, stations such as ‘Sinchon’ and ‘Gangnam’, which are well known, have a wider and more interior than the outer stations(ex: in outskirts).

Second, we can find out lines that are packed with stations of high importance. If there are many stations with particularly high importance on a particular line, we can increase the number of subways running on that line. However, depending on the situation, only some stations on that line (ex: center of Seoul) may be concentrated, and in this case, additional subways that do not go to the end point and only run important stations can be operated for efficience. In fact, there are subways that run only to ‘Sadang’ Station, not to ‘Oido’ Station, which the last stop station of Line 4. This is because there are relatively more people getting on and off near ‘Sadang’ Station(in Seoul) than ‘Oido’ Station(the last stop of Line 4).

Fig10. Example photo of a subway both operating to ‘Oido’ Station, the last stop of Line 4, and a train to ‘Sadang’ Station, the intermediate station

V. Results and Discussion

A. Consider only transfer information (Neighbor nodes; pair of previous-next stations)

df

	index	전철역코드	전철역명	전철명명(영문)	호선	외부코드	다음역	다음역(영문)	위도	경도
0	399	1916	소요산역	Soyosan	1.0	100	동두천	Dongducheon	37.947099	127.060681
1	347	1915	동두천역	Dongducheon	1.0	101	보산	Bosan	37.926664	127.054992
2	421	1914	보산역	Bosan	1.0	102	동두천중앙	Dongducheon jungang	37.914277	127.057158
3	346	1913	동두천중앙역	Dongducheon jungang	1.0	103	지행	Jihaeng	37.901673	127.056409
4	345	1912	지행역	Jihaeng	1.0	104	덕정	Deokjeong	37.889979	127.064305
...	...	...	...	...	...	...	...	...	...	...
451	1002	1716	병점역	Byeongjeom	1.0	P157	세마	Sema	37.206821	127.033268
452	1003	234	신도림역	Sindorim	2.0	234	문래	Mullae	37.50881	126.891206
453	1004	211	성수역	Seongsu	2.0	211	건대입구	Konkuk Univ.	37.544577	127.055991
454	1005	2549	강동역	Gangdong	5.0	548	둔촌동	Dunchon-dong	37.53594	127.132187
455	1006	9999	응암역	Eungam	6.0	999	새절	Saejeol	37.598472	126.915592

456 rows × 10 columns

Fig11. Concatted dataframe of fig1 and fig3

This is data frame after completing (4) of the preprocessing mentioned in III. METHOD above.

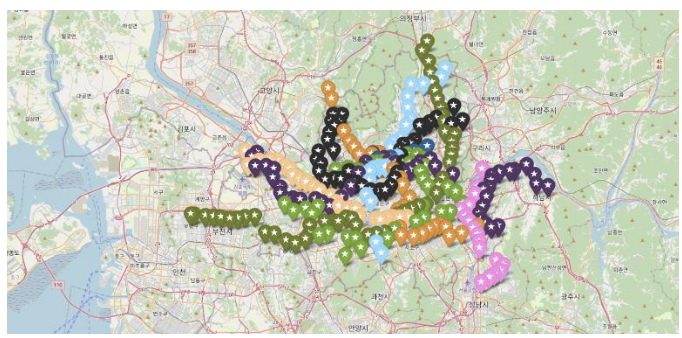


Fig12. Subway stations plotted on a real map

Using package 'Folium' provided by Python, user can plot the location he or she wants to find on Google Maps.We plotted the locations of subway stations on the actual map. We can find out that the latitude and longitude values we specified are correct.

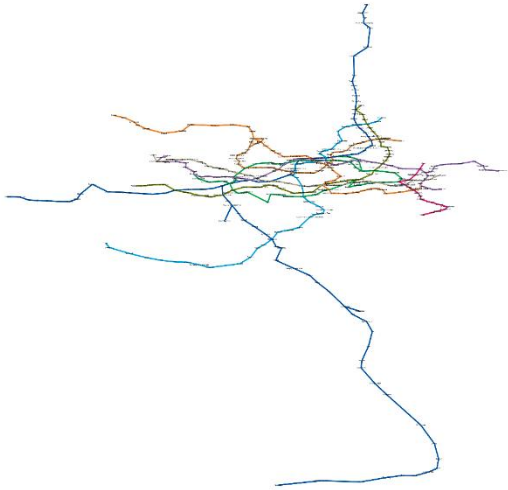


Fig13. Visualization of subway stations by differencing color by line

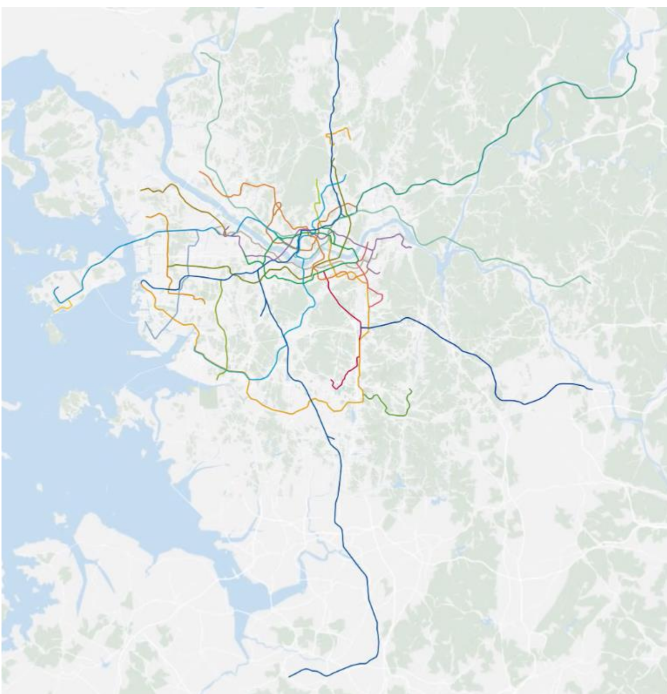
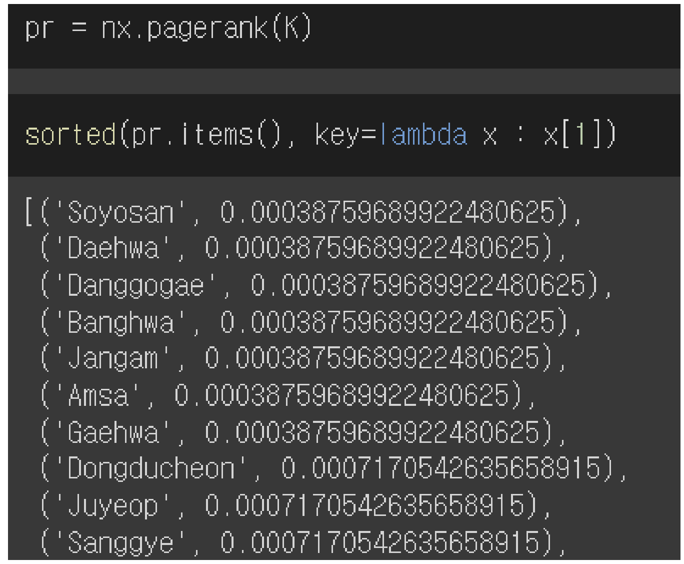


Fig14. actual map containing subway lines

Comparing Figure13 and Figure14, they are similar. So our visualization and data are correct.



```
pr = nx.pagerank(K)

sorted(pr.items(), key=lambda x : x[1])

[('Soyosan', 0.00038759689922480625),
 ('Daehwa', 0.00038759689922480625),
 ('Danggogae', 0.00038759689922480625),
 ('Banghwa', 0.00038759689922480625),
 ('Jangam', 0.00038759689922480625),
 ('Amsa', 0.00038759689922480625),
 ('Gaehwa', 0.00038759689922480625),
 ('Dongducheon', 0.0007170542635658915),
 ('Juyeop', 0.0007170542635658915),
 ('Sanggye', 0.0007170542635658915),
```

Fig15. Pagerank calculation

Fig16. Pagerank visualization

Pagerank is calculated and visualized. But we can find somethings strange.

Figure15 shows in the order of stations where pagerank is high. ‘Soyosan’ and ‘Dangogae’ were ranked high even though they were not familiar to most people. In addition, ‘Sinchon’ and ‘Seoul station’, which are familiar stations to many people, which we often recognize to be of high importance, were even not included in the top 10. Plus, although ‘Gangnam’ was classified very important in the reviewed literature ʻSubway Stations Network Structure Analysis by Using Social Network Analysis’, this is not the case in our results.

Plus, according to our hypothesis, pagerank value of the station in central Seoul, not the outskirts, will be higher. However, in Figure16, there are many areas with high pagerank in stations located outskirts. This does not fit our hypothesis, and in reality, those stations are not recognized by people and are not importantly recognized.

We concluded that the above results were not properly calculated because only too little information was considered. Therefore, additional information that can be more accurate and directly related to importance should be further considered in calculating pagerank. Therefore, we searched more in "서울 열린데이터 광장" and were able to obtain data containing information on the number of people getting on and off by station(Figure 2). We thought that calculating the pagerank by also considering the number of people getting on and off would be more accurate because the actual traffic volume could be considered.

B. Also consider information about people getting on and off

Using weighted PageRank, we can sort the important stations. This is top 10 weighted PageRank results. The stations with high PageRank are same with those we commonly recognize as important, such as Express Bus Terminal and Gangnam. This result implies that our weight estimation is quite significant.

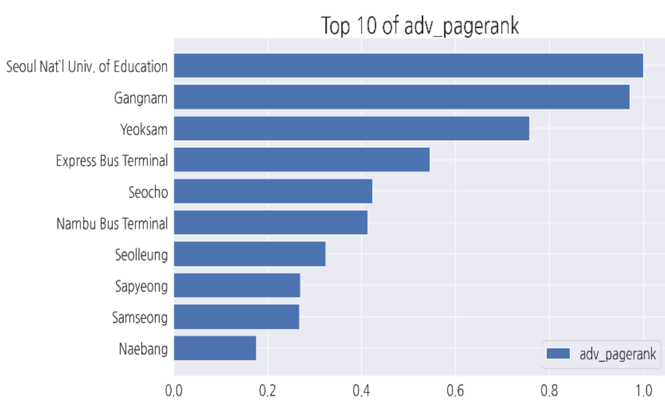


Fig17. Pagerank calculation

C. Additional Analysis

(1) Global Clustering Coefficient

$$C = \frac{number\ of\ closed\ triplets}{number\ of\ all\ triplets} = \frac{3 \times \text{number of triangles}}{number\ of\ all\ triplets} \quad (8)$$

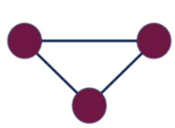


Fig18. Structure of closed triangle

We can derive Global Clustering Coefficient by equation 8. In subway network, there can't be closed triangles because if there are some triangles, it circuits same stations. Since 'number of triangles'=0, global clustering coefficients also becomes 0.

(2) Correlation Matrix

There is very strong correlation between Total # of passengers getting on and off.
There is slightly strong correlation as 0.35 between Total # of passengers getting on/off and weighted PageRank.

Fig19. Correlation Matrix

(3) Linear Regression (OLS)

Fig 14 shows that Weighted PageRank OLS results

OLS Regression Results

Dep. Variable:	adv_pagerank	**R-squared:**	0.131
Model:	OLS	**Adj. R-squared:**	0.122
Method:	Least Squares	**F-statistic:**	14.38
Date:	Sun, 04 Dec 2022	**Prob (F-statistic):**	6.04e-11
Time:	08:09:43	**Log-Likelihood:**	394.33
No. Observations:	387	**AIC:**	-778.7
Df Residuals:	382	**BIC:**	-758.9
Df Model:	4		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	-0.0176	0.016	-1.089	0.277	-0.049	0.014
승차총승객수	3.63e-06	4.94e-07	7.341	0.000	2.66e-06	4.6e-06
경도	0.0615	0.040	1.536	0.125	-0.017	0.140
위도	-0.0245	0.028	-0.870	0.385	-0.080	0.031
호선	0.0002	0.002	0.108	0.914	-0.003	0.004

Omnibus:	526.354	**Durbin-Watson:**	0.258
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	58418.229
Skew:	6.785	**Prob(JB):**	0.00
Kurtosis:	61.641	**Cond. No.**	1.39e+05

Fig20. PageRank OLS result

Only p-value of number of passengers are less than 0,05 It means that # of passengers is only significant.

VI. CONCLUSION

We visualized the metropolitan subway network using data on the relationship between subway stations in the metropolitan area and the data on getting on and off, and analysed it by calculating the PageRank. As a result of the Literature review, Gangnam Station and Jamsil Station showed high importance or centrality, and it is also highly recognized by most people. We conducted a task after establishing a hypothesis that the center of Seoul would be of high importance and that the transfer station would be of high importance. When PageRank was calculated using only pair information with surrounding stations, the results were different from both the Literature review and the hypothesis.

To solve this problem, we additionally estimated the weight using the data of the number of people getting on and off the subway. Using the estimated weight, the weighted PageRank was calculated using the maximum eigenvalue method. The top 10 PageRank included most of the stations frequented by people such as Express Bus Terminal and Gangnam. This means that the weight we estimated is appropriate and that the weight is well done. In addition, as a result of correlation analysis and OLS analysis, it was found that the weighted PageRank was somewhat proportional to the number of passenger's boarding and exiting.

REFERENCES

[1] Kyung Hoon Han, Su Seo, Geun Byung Park, “사회 네트워크 분석을 이용한 지하철역 네트워크 구조 분석”, 2015 년도 한국철도학회 춘계학술대회 논문집, no. 8, pp. 1-7, 2015

[2] 서울 교통공사, “서울교통공사 노선별 지하철역 정보”, 2018.11.08., 서울 열린데이터 광장, https://data.seoul.go.kr/dataList/OA-15442/S/1/datasetView.do, 2022.11.28

[3] 서울 교통공사, “서울교통공사 지하철 환승역 환승인원 정보”, 2015.02.27., 서울 열린데이터 광장, http://115.84.165.39/dataList/OA-12033/S/1/datasetView.do, 2022.12.01

[4] Metro Liner, 2021. 6. 29., “지도로 보는 서울 수도권 전철 1974-2021”, https://www.youtube.com/watch?v=eoyW-BMNnqg, 2022.12.01

[5] *Using Complex Numbers in Website Ranking Calculations: A Non-ad Hoc Alternative to Google’s PageRank - Keita Sugihara* https://dl.acm.org/doi/abs/10.1145/775152.775190