Application of Graph-based Network Analysis to Champaign-Urbana Transit Networks

Keywords: Graph, Network Analysis, Transit Network, Preferential Attachment, Nodes, Edges, CUMTD

Abstract
This paper explores the application of graph-based network analysis and descriptors to a real-life transit network. Transit networks are typically studied within the geographical context that they are situated in. Using a graph-based method allows us to better explore questions of importance and connectivity in an intuitive manner. Our research question is: what new information can we learn by studying transit networks devoid of a geographical context? We compare the effectiveness of various ways of visualizing and studying the Champaign-Urbana transit network. We started with the motivation that our project will help to identify popular places and important transfer locations in town. Our assumption is that it will be useful to get a better idea of the city, especially for new international students. From our results, we found top most popular spots by degree, betweenness, closeness, and eigenvector in the city. Surprisingly, the identified stops are not often discussed among students as popular stops. Hence, it shows that our study helped to identify some new and useful information.

Introduction
We set out to examine new information that can be learned by studying transit networks devoid of a geographical context. This is very useful because we might not be able to access geographical information all the time. Our assumption is that it will be helpful to understand the transit network in Champaign-Urbana from a different perspective without geographical context especially for people new to the city who are not very familiar with different locations in town. The Champaign-Urbana Mass Transit District (CUMTD) is a network of bus service serving Champaign Urbana and its environ.

Our research aims to compare the effectiveness of various ways of visualizing and studying the CUMTD network. Because the University of Illinois community receives several international students every year, our focus is to develop a way the transit network can be easily visualized by new students to quickly identify the important locations around town. Using our results, we expect that it will be useful to get a better idea of the campus, especially for new students. Our results also provided information about the most popular stops on campus based on node degrees. For new students, it would be good to have a list of places to visit on campus at the beginning to be familiar with campus and its most accessible and popular spots. It would also be safe to know more places on campus mainly for international students who have come from a variety of backgrounds and locations.
Our study allows us to potentially draw different conclusions from a different approach to the study of a traffic network. This could have implications for transit planning and other urban planning related work. Our assumption is that a different way of visualizing road network will help users identify where popular places are, and where important transfer locations are. To help with our analysis, we first collected MTD bus data online, processed the data for analysis using Python, and we drew our conclusion based on the result of our analysis. The rest of this paper expands on these introductory ideas.

Background

Public transportation plays an important role in people's lives in all urban areas. As a result, many attentions have been drawn to optimization of transit networks. The benefits of graph theory have been widely illustrated by several researchers, and it is agreed that a selection of indicators will be needed to measure and help with the understanding of a transit network (Fortin, Morenci, Trepanier, 2016).

Derrible and Kennedy (2011) adopted the methodology of visualizing transit network into graphs, based on greater Vancouver regional district public transportation system.With the transit network described in a graph, the properties of the network can be computed, city planners and practitioners could get a better idea of vertices and edges and their relationship in the graph. Also, in Derrible and Kennedy's research, improved network connectivity indicators were proposed along with directness, degree of complexity, and these make a good measure for future bus network study. A similar study by Quintero-Cano, Wahba, and Sayed (2014) introduced a new procedure for drawing bus networks as graphs, by disaggregating them into sub-networks at the traffic analysis zone level. Quintero-Cano et al. analyzed the effect of bus route transfers by introducing intermediate walking transfer edges. They also proposed improved network connectivity that incorporated the influence of bus operational characteristics.

Real world networks have been observed to follow certain patterns. In the extant literature, these patterns have been referred to as the Matthew effect, scale-free networks, power law, or preferential attachment. This phenomenon has also been studied in the urban transit domain. Wu, Gao, Sun, and Huang (2004) studying the urban transit system in Beijing, China, concluded that the transit network in the city is a scale-free network based on the outcome of their model. According to their study, growth and preferential attachment help the urban transit system develops into a scale-free network. They arrived at this conclusion by conducting experimental studies using transit data from Beijing.

Their study shows that for an urban area, the development of the transit system follows a scale-free network approach. According to their explanation, this is mostly because as the urban area continues to experience growth, transit routes are added based on ongoing development and most of these usually favor existing transit system. In other words, roads are built to link existing roads and through this, a scale-free network emerges. Wu et al. (2004) also note that the scale-free network has a serious drawback. This drawback is its vulnerability to attacks. Their paper considered only non-stop urban transit networks and neglect the effects of middle stops between a node of origin and the destination node.

Data

This data is derived from data prepared by CUMTD for use in their General Transit Feed Specification (GTFS), which is developed by Google to provide transit directions in Google Maps. This data was downloaded on 3/11/18, and all results from this study should be treated accurate as at that date. Changes that are made after this date will not be accounted for. Since this data is prepared by CUMTD for Google Maps, we believe that this is the most reliable data with which we can work on. The data can be found here: https://developer.cumtd.com/gtfs/google_transit.zip.

We chose to use this data as it has data about every single scheduled run of every route in the CUMTD network, along with every stop. This is great for processing into a simpler data set that can give us node and edge data depending on how we want to apply graph theory to the network. From GTFS, we constructed a table of node pairs with weighted edges using R and Python for preprocessing. R was used to grab columns of interest from the separate files of GTFS and combine them into a single table for processing. Python was then used for further preprocessing as loops are more easily constructed in Python. Also, making the transition to Python at this step allows us to keep things simpler when we begin utilizing the NetworkX package.

Edges and Nodes

We made a number of assumptions in the process of collapsing our transit network into a graph. In our case, nodes represent the 1117 bus stops in the CUMTD network, and edges represent bus services. We assume that if two stops had a bus service between them, they will have an edge between them. This collapses the graph to be entirely void of any sort of geographical context. Stops that are further down the route from a given starting stop are represented as having a path length of 1. This assumption allows us to characterize what are the most well-connected places on the network, as that would be the bus stops with the highest weighted node degrees. Node degrees, in this case, would represent the number of bus stops a person could reach by taking a single bus without transfers. This assumption further reduces stops along a given route as a complete graph of edges weight 1.

For data simplicity, we also made the assumption that all stops in the network serve bidirectional routes. This is a big assumption that we are making, as there are obvious exceptions. However, these exceptions typically have buses running in the opposite direction pass on a set of one-way streets nearby, and therefore can be considered as serving the same area. For this project, we picked the segments of routes indicated in GTFS with *direction_id=1*.

Lastly, we treated variations of a service as unique (i.e., 5W Green and 50W Green Weekend) and treated similarly in weight. This assumption is valid as a bus stop that is being served by multiple variants of a route is likely to be very important in the network, either serving an area where demand is consistently high across all periods of the week or serving as a main transfer point for the network. This assumption comes around because we used data from GTFS for all available services, and did not account for variation in service by time of day or day of week.

Method

Network research methods are composed of three parts, from network construction to building metrics based on nodes and metrics based on graphs. Several centrality measures are adopted in the analysis, among them, degree, betweenness, closeness, and eigenvector centrality are most valued. With the undirected and weighted network built, the Average Path Length, component coefficient, density and so on are also considered from the viewpoint of a whole graph. With the definition of nodes and edges, the network is one giant connected components, but since nodes are well connected together which makes the visualization hard to show any insightful ideas, another method is taken to emphasize some really important bus stop. The edges whose weight are less than 10 are filtered out of the later analysis, so the whole network is broken down into smaller cluster with its center bus stops. Among those divided clusters, the largest connected component is further analyzed, and along with the whole network, their weighted node degree distribution is plotted for analysis of their network type.

Results

With the help of four centrality measures, top ten nodes are reported in Table 1. Among these four measures, we could find some bus stops occur under more than one measures. This shows the influence of such bus stops in the traffic network. The higher they are in the ranking, the popular or important they are. For example, bus stop 'Round Barn Road' is high in the ranking of three centrality measures, it turns out to be located in downtown Champaign and has more than 10 buses stopping there. Illinois Terminal (platform B) is another bus stop that plays an important role in the bus network. It is a transit stop for many bus routes. It appears this table uncovers much truth about the bus system of Champaign-Urbana district from the perspective of network analysis.

Table 1. Top Ten Bus Stops in Four Centrality Measures

Node	**Degree**	**Node**	**Betweenness**
Round Barn Road	0.5436	Urbana Middle School	0.0325
Illinois Terminal (platform B)	0.4905	Crescent and William	0.0278
Springfield & country Fair Dr. (SW)	0.4714	Edison Middle school	0.0213
Green & Goodwin (SW Corner)	0.4514	Bradley and Clayton	0.0208
Bradley & Country Fair Dr. (NE)	0.4313	Parkland College	0.0188
Bradly and Clayton	0.4273	Green & Goodwin (SW Corner)	0.0186
Walnut & University (SE Corner)	0.4253	Central High	0.0183
Church & Neil (SW Corner)	0.4253	Market Place North Entrance	0.0175
Walnut and Logan	0.4253	Fourth and Peabody	0.0168
Main & Fremont (South Side)	0.4223	Goodwin and Nevada	0.0167
Node	**Closeness**	**Node**	**Eigenvector**
Round Barn Road	0.6866	Green & Goodwin (SW Corner)	0.1417
Illinois Terminal (platform B)	0.6624	Round Barn Road	0.1403
Springfield & country Fair Dr. (SW)	0.6542	Illinois Terminal (platform B)	0.1356
Green & Goodwin (SW Corner)	0.6403	Main & Fremont (South Side)	0.1337
Bradley & Country Fair Dr. (NE)	0.6358	Walnut and Logan	0.1332
Bradly and Clayton	0.6358	Church & Neil (SW Corner)	0.1332
Walnut and Logan	0.6350	Walnut & University (SE Corner)	0.1332
Walnut & University (SE Corner)	0.6350	Mattis & Round Barn (SW Corner)	0.1300
Church & Neil (SW Corner)	0.6350	Springfield &Mattis (SW Corner)	0.1300
Main & Fremont (South Side)	0.6338	Goodwin at Ceramics Building	0.1200

As a fully connected network, the metrics for the whole graph also need to be considered. The average path length of the network is roughly 2. The interpretation here is that the average number of bus people need to transit to any nodes in the network is two. The clustering coefficient is 0.7747 which is relatively high. These two values suggest the bus network is well-built and successful one, people can easily get to their destination without much effort.

Table 2. Additional Graph Metrics for the Traffic Network

Average Path Length	2.0275
Clustering Coefficients	0.7747
Density	0.1341
Transitivity	0.5715
Connected Components	1.0000
Diameter	4.0000
Radius	2.0000

Since this is a network, another interesting question will be: what is model under this network? To figure out this question, the distribution of the weighted node degree is drawn below in Figure 1. A power law trend can be detected which is the feature of a scale-free network, combined with our knowledge of bus network and the visualization of the largest connected component in Figure 2, we can safely say the bus system in Champaign Urbana district is a scale-free network model.

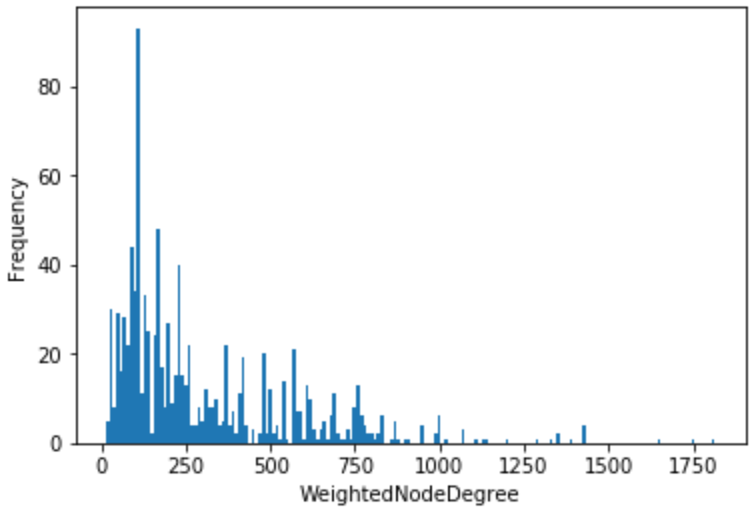

Figure 1. Node degree distribution of the network

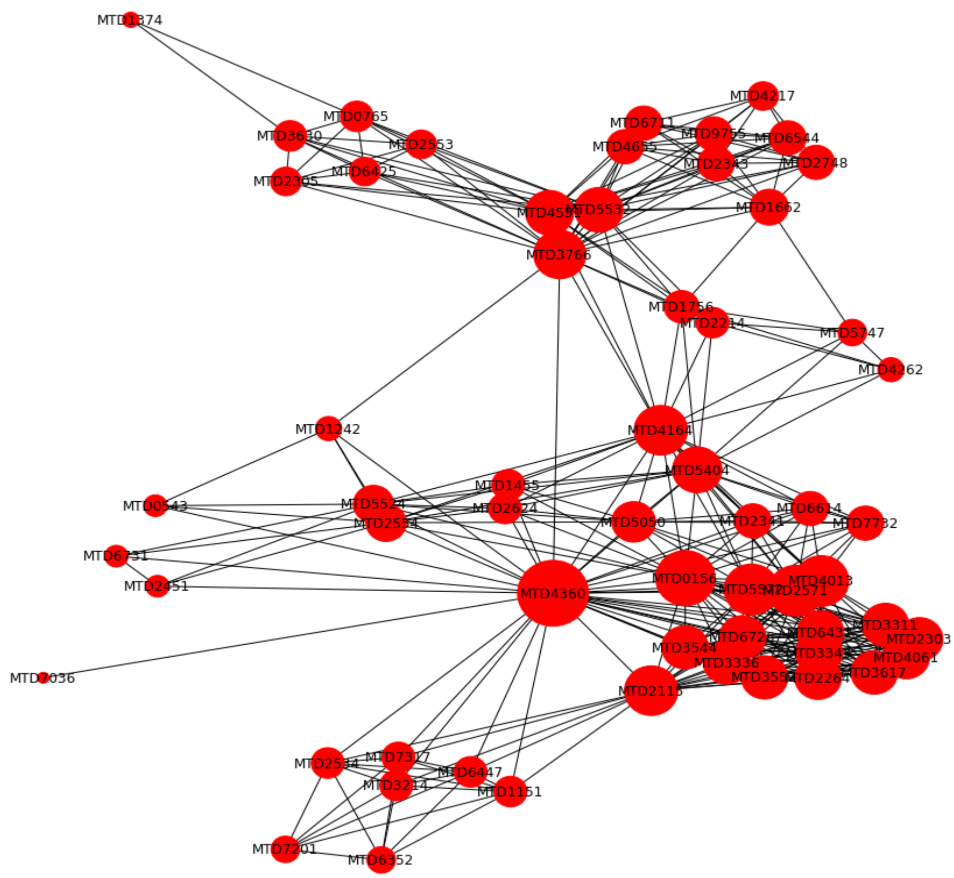

Figure 2. Largest connected component

Conclusions

Our main conclusion from this project is that we can get useful information by studying transit network devoid of geographical context. This is very interesting knowledge since people or students who are new to the city or the university usually need access or try to understand geographical locations during their first few weeks. We also found that our results are matching with the previous research hence these could be generalized. Our results are most useful to the new international students at the University.

Limitations and Future Work

Through this project, we have reduced the entire CUMTD network into a graph with assumptions that are atypical to a transit study. Conducting the study on the entire network can be a limitation to its usefulness, as transit planners tend to think of transit services by time of day and day of the week (i.e. weekday morning peak, weekend non-peak, etc.) which is why CUMTD has variations in their routes. A similar study accounting for these temporal variations could be carried out and may provide greater insights into the network. At this level, all we can really say from our analysis is how important a stop is in the entire network.

Furthermore, the way we have characterized edges removes an important piece of a transit network. Notably, distance is abstracted out of the study. Our graph cannot tell you how far apart 2 stops are, but we fell that this limitation is mitigated by the ability of us to access the number of transfers required between stops, which is an important metric when it comes to transit user comfort. Treating variations of a service as unique generated interesting problems in our graph.

For example, path length, which we thought to be transfers, seem to suggest something else completely. We think that this is because of how there are many services out there that run very low-volume routes. Some routes are only available at certain times of the day and week, resulting in them being registered as an additional transfer away when that transfer cannot be realistically taken. It may, therefore, be improper to apply our analysis to the entire transit network, and further studies can attempt to apply the same measures on a temporal subset of the network. Our study can further be extended to consider directionality and sequence in a route. Removing geographical context seems to produce results which usefulness is limited by this very fact. Directionality, sequence, and temporal dimensions are all important parts of a transit network and should be considered.

References

Derrible, S., & Kennedy, C. (2011). Applications of Graph Theory and Network Science to Transit Network Design. Transport Reviews, 31(4), 495–519. https://doi.org/10.1080/01441647.2010.543709

Fortin, P., Morency, C., & Trépanier, M. (2016). Innovative GTFS Data Application for Transit Network Analysis Using a Graph-Oriented Method. Journal of Public Transportation, 19(4). https://doi.org/http://doi.org/10.5038/2375-0901.19.4.2

Quintero-Cano, L., Wahba, M., & Sayed, T. (2014). Bus networks as graphs: new connectivity indicators with operational characteristics. Canadian Journal of Civil Engineering, 41(9), 788–799. https://doi.org/10.1139/cjce-2014-0054

Wu, J., Gao, Z., Sun, H., & Huang, H. (2004). Urban transit system as a scale-free network. Modern Physics Letters B, 18(19n20), 1043–1049. https://doi.org/10.1142/S021798490400758X