Mathematical model for cost-efficient installation of public transportation system

Eui Seok Kim[*]

University of California at Berkeley
digimom321@Berkeley.edu

Abstract

This thesis analyses the way how theorems in network theory could be applied to choosing locations for placing each of bicycle stations in the city efficiently. Hence, it sets up a mathematical model for predicting paths and number of people interacting at the given time interval using graph theory and game theoretical techniques. Then, it evaluates the cost efficiency of such system and concludes with case studies to show how the model could be used in real life setting.

I. Introduction

As there has been a rise in concern for issues regarding congested transportation system in urban settings, more and more cities have been developing systematic public transportation system. When it comes to adopting a whole new system of transportation system, which involves various steps of setting up a complex network, it is tremendously important to consider efficiency. There could be many different ways to define efficiency and plenty of mathematical theorems from field of combinatorics that could be used to find the optimal circumstances for the installation of such system.

In particular, this thesis concentrates on how to increase efficiency of the system via choosing optimal locations for transportation stations by adopting graph-network theorems and other references from combinatorics. To be more specific, by applying these theorems, this thesis models formulas that give anticipated number of people moving from certain point of location to the other. Hence, based on the graph theorems, it calculates the required number of bicycles and stations at each potential location.

This process involves some underlying assumptions and definitions of potentially ambiguous terms that will be mentioned significantly in the thesis.

II. Backgrounds and Assumptions

As mentioned in the introduction, how to install optimal system for a public transportation is a broad question. Hence, it is very ambiguous since it would involve a lot of different parameters that might affect efficiency, depending on how we define. However, through the analysis of some already existing modern public urban transportation systems, we could narrow down complexity with few assumptions. Also, from now, let's define this transportation system as a system that consists of bicycles and bicycle stations located at each key sites.

On transportation users

[*] A thank you or further information

- In the theorems and models, we only count the number of people who are mobile(always moving during each time interval ti to ti+1 where $0 \leq i \leq n-1$ and there are n number of time intervals)
- Everyone residing in city is willing to take advantage of the system.
- Every rider move same amount of distance per each time unit.
- Every group of riders move in the same pattern during the weekdays(Mon to Fri) and Weekends(Sat and Sun)
- Every ride costs the same amount of money, and the marginal satisfaction regarding the price is same for everyone.
- There are groups of riders who start their day in the same station and return and end their day in the same station.

Now we have narrowed down to a system which contains users with uniform characteristics, same marginal benefit and uniform moving pattern. Hence under these assumptions, it is now our job to figure out during which time interval how many people will move from arbitrary locations A to B. Since now, we call the movement a 'flow of people.'

Before, we delve into mathematical proof of getting models for expected number of flow, we have to set up another list of assumptions for the system itself.

On the system

- Every station contains some bikes and some ports.
- Every two stations are close enough that people could move from arbitrary station A to B in a single time interval.
- Every bike is capable of moving the same distance in certain amount of time.
- Every station is operated in first come first serve system.
- Every station allows bike rental after it receives sufficient number of returned bikes at each time interval.
- In every station, bikes are uniform and could be parked in every other station.

This completes basic fundamental assumptions and build-ups for our model. Hence, even though it might be very confusing, we would have a better understanding of this model as we get through some figures and mathematical theorems that help to explain this model.

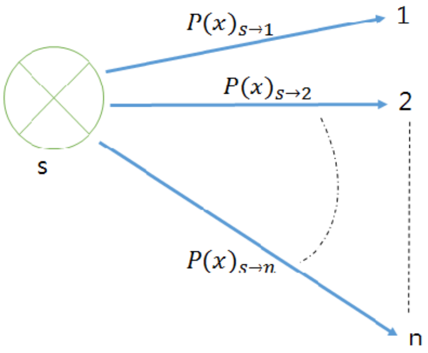
III. Construction of a model

As we have defined in the above section, the transportation system is automatically operated, but driven by decision makers who rent out bicycles and return them into the stations. However, the main concern and tricky part here is that we have to come up with organized structure and elaborate mathematical function to keep track of movement of people. Hence, the number of people moving from arbitrary location to another at arbitrary time state is the key element that decides the number of bicycles and ports needed at each of the station. It would be very useful to adopt graph theory to locate each station and track the flow of people.

From now on, we simplify our model by the notation (V,E;T) where (V,E) denotes the graph G where V represents the vertices, or the bicycle stations in this model and E stands for the edge of the graph, which is equivalent to movement of people from one station to another. Then T represents the set of time steps, during which people move from points to points using bicycles. Hence, we set up another set of functions P, which denotes the weighted value of number of people who are moving from point to point. This P value is then written as the degree of each edge. To decide where to locate each vertices, we have several different criterion. Please wait until I get back to it on **Section IV, subsection I**

In order to construct the set of functions, denoted P, to track the tendency of movement of people, we first have to measure the variation of the degree of each nodes of (V,E) in respond to time. Suppose that we have a graph (V,E) described as follow figure . Hence consider the set of time measures $T = \{t0, t1, ..., tn\}$ where whole active hours suppose for the arbitrary city, 8:00am to 8:00pm is partitioned into n number of intervals $\{t1, t2, t2, t3, ..., tn - 1, tn\}$

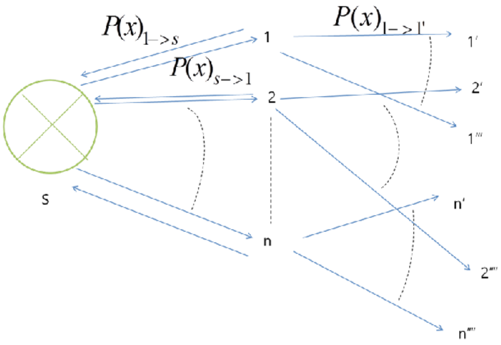
Figure 1: *A movement tracked from time t0 to t1.*



Suppose that people start off their day at time t0 at some arbitrary starting train station S and start to ride bicycles from S to other vertices 1,2,...n. Then let's let the function P(x)s->1 describe the amount of people moving from point s to 1 in some time interval. As you see in the figure 1, every edge of the graph (V,E) describes the total number of people moving along the edge.

Hence, we assign the sign value to each of P(x)i->j. This is such that if ∣P(x)j->i∣=∣P(x)i->j∣=k, then P(x)i->j=k and P(x)j->i=-k. This represents the weighted value from the number of people moving from i to j and j to i, respectively. With this definition we can assign P(x) value for every single pair of points (i,j) in each time step ti.

Figure 2: *A movement tracked from time t1 to t2.*



Just to make sure, don't forget that there could also be P(x) function defined between any two points of 1,...n *I am omitting these in the figure just to show clear difference between (V,E;t1) and (V,E;t2).* Now, take a look at Figure 2,the graph at the second time step t2. We see that compared to the graph structure (V,E;t1), the graph at time t2, denoted as (V,E;t2), has more edges and vertices connecting and connected to each other. Hence, more importantly, there are edges with both P(x)i->j and P(x)j->i are non-zero where i and j denote some vertices in the graph. Also, from points $\{1, ..., n\}$, there is extra edges connecting these points to the corrsponding points $\{1', ..1'''(itimes), ..., n, ..., n'''(ktimes)\}$. where i and k are intergers larger or equal to 0.

Note that the graph of these edges and vertices as described in the figures above is only a single part of the whole graph (V,E;T). We may assume that there are finitely many parts of the graph that shows similar patterns of directions and weighted values of edges, and all those parts are combined together to form complete picture of (V,E;T). From now on to later in the paper, we may call this "preview" version of our graph as (V',E';T). One another important thing to note is the graph (V,E;t2) has the opposite directional edge P(x)1->s, but it does not contain P(x)1'->1, and we know that it is impossible because people are only capable of transporting a single edge step per single time step. Hence, we can extend (V',E',;T) inductively respect to the time step tn.

Now, in the following section, we would want to look at the way how to optimize the system in order to suite transportation users under the assumption. This can be limited to deciding number of bicycles and ports in each station. Hence, there are several different perspectives of defining the efficiency, or so called 'optimality.'

IV. Optimality

Building an 'efficient' model depends a lot on how we actually define 'efficiency.' There might be several ways of defining it. Note that in the first section, we have assumed that cost of bicycles are fixed and every people in the city are parts of the bicycle transportation system. To meet this criterion, we would first define being 'efficient' as being 'able to serve everyone who needs to be part of the system." This definition directly affects the minimum number of bicycles and ports that each station has to include.

Definition 1. *A system (V,E;T) is* ***efficient*** *iff it contains the minimum number of bicycles and ports that could satisfy every participants*

The most important thing to note in this definition is that it is well-defined based on the fact that whether the system 'satisfies' participants or not. Let's suppose from now on, by 'satisfaction,' it means the system has enough number of bicycles and ports to keep it running without anyone having to wait until he/she gets the bike.

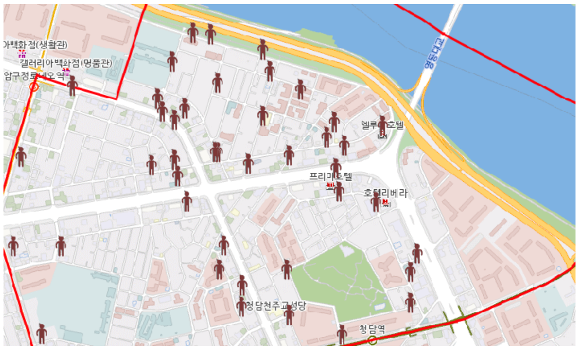
Now on behalf of this definition, we will try to find the optimal location of station, number of bicycles and ports at each station, respectively.

I. location of stations

As mentioned in the beginning of the chapter, each station corresponds to each of vertices in the graph (V,E;T). In choosing where to locate vertices on the actual geographic map of the city, in other words, on deciding where to install bicycle stations, The population that passes through the each candidate points matters a lot. More specifically, as there are more people passing certain candidate points, those points are likely to be the places where stations would be located. Let's take for example, the city of Seoul. Below Figure 3 is the data of float population at each time at each points.

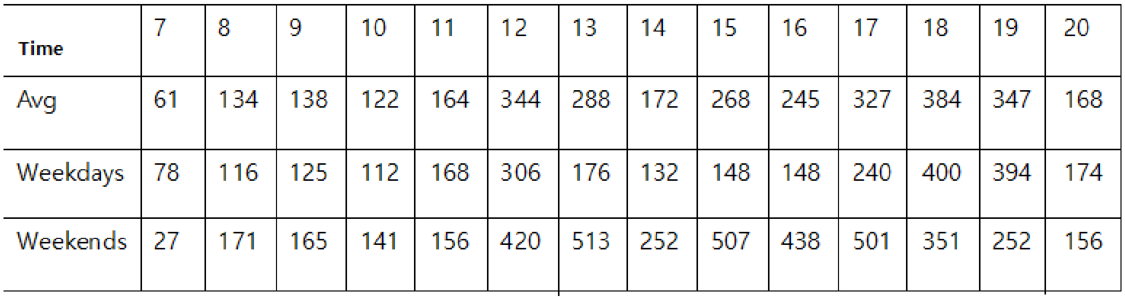
In this map, the point with the person icon on it contains the data of number of people at passing by that certain location in each time step. Hence, in the selection process for choosing 'where to investigate,' we set up a lower bound in terms of average float population in each place and get rid of candidate points that do not satisfy this lower bound criterion.

Figure 3: *A map with the data of float population of specific points at each time step.*



More specific data of float population can be viewed when we click on the icon at each location. The chart below provides the average value of float populations in each time step (7:00 am to 8:00 pm) of weekdays and weekends, respectively. Note that these data was gathered by volunteers standing at each point, counting numbers of people who were passing by at specific time.

Figure 4: *An example of data of float population of Plaza hotel at Chungdam city, Seoul in each time step, containing overall Average, Weekdays average, and Weekends average, respectively.*



Time	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Avg	61	134	138	122	164	344	288	172	268	245	327	384	347	168
Weekdays	78	116	125	112	168	306	176	132	148	148	240	400	394	174
Weekends	27	171	165	141	156	420	513	252	507	438	501	351	252	156

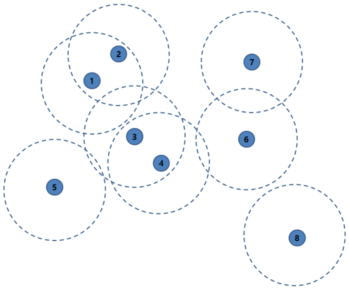
By the first definition of efficiency, in order to suite more people, it is very clear that we have to pick at least the points where the most number of people pass by on average based on the data. After deciding the number n of bicycle stations that we are going to install, the process which is determined by more game theoretic approach that will be introduced briefly toward the end of the paper, we choose n number of points with the most number of average float population. Suppose now we have a set of vertices $\{1, ..., n\}$, of which each of the element would be locations for bicycle stations.

Now, we have chosen fundamental building blocks of E of our model (V,E;T). By the assumption in the very first and second pages of this thesis, we know that people can move at most same limited amount of distance using bicycles during a single time step ti to ti+1. Suppose we have a map with only the vertices 1,...n, which are located n accordance with the real geographical distance. We call this the "range." The definition is as follows.

Definition 2. *A point i is in the range of another distinct j and vice versa iff a person at point i can reach j or a person at j can reach i in a single time step.*

suppose we have chosen 8 candidate points. Let's denote them $\{1,2,3,4,5,6,7,8\}$. Hence, assume that the actual geometric locations are as follow. Here, note that each small circles denote those eight chosen points and large circles denote the range of each point that we have defined above. The hence in building a full picture of graph (V,E;T), we follow algorithmic steps.

Figure 5: *Arrangement of 8 candidate vertices according to real geometric positions.*

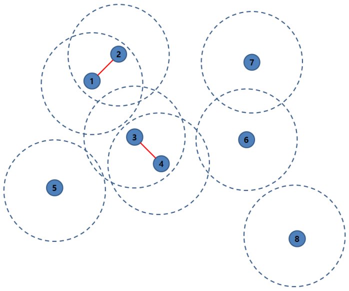


From here, let us define sets $R1 = \{1,2\}, R2 = \{3,4\}$ and $SemR1 = \{1,2,3\}, SemR2 = \{3,5\}, SemR3 = \{4,6\}, SemR4 = \{6,7\}$. As you can easily infer from the figure above, set Ri refers to the the set of vertices that are in the range of each other, and i is just an index. Set SemR refers to the 'semi-range,' and also i is an index. The definition of 'semi-range' is as follows.

Definition 3. *Points i,...,j are in the semi-range of each other iff every one of their ranges intersects every one another, but not every one of them are in the range of each other.*

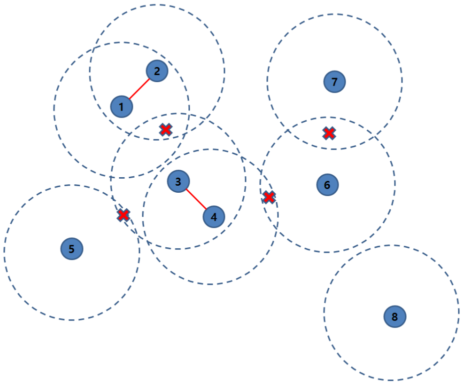
The next step is to connect points in the set R1 and R2, and then we see that the edge correspond to possible route for bicycle rider to move during one time step.

Figure 6: *Every points in the sets Ri are connected.*



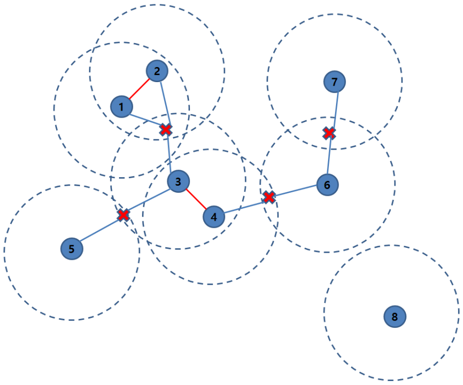
After connecting those vertices that are within a single time step distance, we might realize that we are far away from getting a full graph (V,E;T) because we do not have a way to measure the population of the people who move from a point to the other point that is more than or equal to 2 time step away distanced. In order to have more solid model, we proceed to add one vertice to each of Semi-range sets. Hence this results as below

Figure 7: *x denotes the extra point that was added to each of SemRi sets.*



The intention behind this step is very simple. I the very first section, we have assumed that every individual participant of the system moves in the same speed and same unit per single time step. The problem might arise when we try to track the number of people who move toward the points that are more than or equal to 2time steps away from where they are standing at currently. Suppose that some descent number of people larger than 0 are moving from point 1 to 3 during time step ti to ti+2, in the above example. Since, the vertices 1 and 3 are not connected to each other, P(x) value coming out from point 1 during the time step ti to ti+1 and coming in to 3 during time step ti+1 to ti+2 cannot be evaluated, and there appear mismatch in the data.

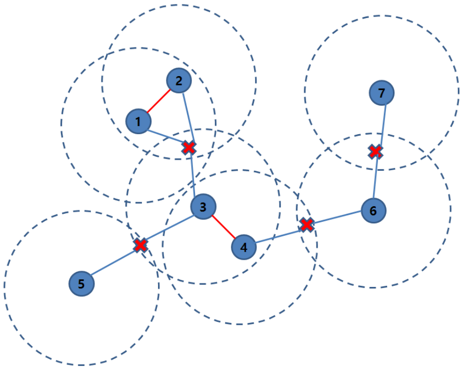
Figure 8: *Include each x as the part of the elements of set E, the vertices of full graph (V,E;T), and connect every adjacent pair of edges*



After drawing extra points on the map, we have the complete set of the vertices E by now. Henc,e we proceed b connecting every adjacent vertices of E. By 'adjacent,' I mean that two edges are within the single time step distance.Hence, we get a complete graph above.

The last step involves only a single effort. Most of the times it would be the case that we face the situation where there exists one or more points which are in neither of range and semi-range of other points. In this case, we just simply get rid of this point from the set E, and possibly try to include it in another set of graph. Now, we obtain full picture of a graph system (V,E) as follows.

Figure 9: *Now we have a complete set of the graph (V,E)*



we repeat this algorithmic steps inductively until we run out of adjacent vertices that could connect to the endpoint of the graph. Hence, we only consider the case which among the initially chosen points, there exists at least one non-empty Ri set and at least one non-empty SemRi set. Hence, we only consider where (V,E) is a conected graph.

Theorem 1. *If the graph (V′,E′) is a connected graph with three or more vertices, then it is the part(subgroup) of some full graph system (V,E;T)*

Proof Suppose that the graph (V′,E′) is connected but not a part(subgraph) of another graph system (V,E). Then it is either i) parts of multiple different graph systems, or ii) it contains extra edges and vertices that are not part of the system (V,E). In case i), it is impossible since in the first step of the algorithmic construction, some vertices that are in different graph systems would have been matched and integrated into full system later in the algorithmic process. In case of ii), it is also impossible because , again by the algorithmic construction, some vertices that are not in (V,E) and some vertices in (V,E) must have been elements of the same set Ri or SemRi.

Some might say that this theorem is simply just a corollary of the definition and algorithmic creation of the system (V,E;T). However, we might find this extremely useful in the later in the paper when we want to expect what would happen on the full system (V,E;T) by observing much simpler system. Also, when we are considering a system with infinite number of vertices and edges, we can simply use Compactness property of trees and graph and try to find finite subgraph that could represent similar behaviors.

II. number of bicycles

By what we have defined and proven in the previous section, the time has come to actually start taking a closer look at each bicycle station in each vertice. Hence, the number of bicycles and ports at each station is the most fundamental determinants that decide how much the system (V,E;T) meets the criterion of 'efficiency,' which we have defined in the first definition. Now to remind the readers of what we have discussed in the first chapter, the brief structure of bicycle station is consisted of bicycles and ports, and when bicycles are rented out the space where they were originally parked becomes ports. Hence, people can only return their bicycles when there are available ports and have to wait when either there are not enough ports to park bicycles or there are not enough bicycles to rent them out.

Consider the Figure 9. and Figure 1..Hence keep in mind that Figure 9 and Figure 1 are both full system of the bikes, but Figure 9 is undirected and not yet contain the data in regards to the population movement function P(x). These two graph formats seem totally different, but we can go through weighting of the edges and make them graph homomorphic to each other. From what we have got in Figure 9., apply P(x) value to each of the edge and each of the direction. Then, set up a starting point of the model. Denote that vertice as 'S' and position in the front, just like how it is in Figure 1. Then, right next to S, place elements of E that are in the range of S in the same column. Then, in the next column, try finding the elements of E that are in range of points i's, which were the vertices that were in range of S. Repeat this process inductively until we run out of vertices in E. Now the undirected and unweighted graph of form Figure 9. has been homomorphically transformed into full system (V,E;T).

Hence, we may have a theorem on number of bicycles.

Right before going into theorems Define Mt1(X) as the summation of all of P(x) values coming in and going out from a vertice x to adjacent edges during t0 to t1. Similarly define, Mt2(X) as the sum of all of P(x) values coming in and going out from a vertice to adjacent edges(note that by the definition,if we assign P(x)x->i positive value or 0, then P(x)i->x is assigned negative value or 0, and in this defintion of M, P(x) value of population coming in to point X is less than or equal to 0 and of population going out from point X is larger then or equal to 0)Hence, more generally, define Mti as summation of P(x) value in time step ti-1 to ti.
More formally, we write Mti(X)= Pti(X->1)+Pti(X->2)+... + Pti(X->n)+Pti(1->X)+...+Pti(n->X)

Now we have a theorem.

Theorem 2. *The number of bicycles that satisfies the definition of 'efficiency' in Definition 1. equals to Max[Mt1(X), Mt2(X), ..., Mtn(X)].*

Proof we see that there are three cases possible. i) when Mti(X) equals to 0 ii) when Mti(X) is positive iii) When Mti(X) is negative.
In case i), it is either Mti(X) contains 0 summands(there is no demand and supply of bicycles at all) or Mti(X)has more than one summands, but it equals to 0. (if follows from the definition of P(x) that there are same number of people coming into and out from point X) Now, consider the second case ii). When Mti(X) is positive, it means that there are more people going out from point X to other adjacent vertices than the people who are coming into point X from other adjacent vertices. This, by the definition of the function P(x), directly implies that there are more people who need to rent out bicycles than the people who need to return their bicycles and park them. Lastly, consider the case iii), using similar arguments. In this case, Mti(X) being negative implies that there are more ports needed than the bicycles since there are more people who are returning their bicycles than the people who are trying to rent out bicycles. Hence, by the construction and assumption that every bike renter has to wait until every bike returner returns their bikes. Suppose that we leave the bicycle station without any bike parked originally. Then we see that in case ii), Mti(X) equals to the number extra bicycles needed to suite everyone who has waited for all the bikes to return but did not get anything to rent out at time step ti-1. Hence it is the number of bicycles that coincides the definition of "efficiency," defined in Definition 1. By taking the maximum among the values of it in different time steps, it equals to the minimum number of bicycles to be placed on time t0 in order to suite everyone in time every single time step. Hence, it

suffices the definition of "efficiency," defined in Definition 1.

We see that by the Theorem 2., putting the right amount of bicycles at each station at the very first time step enables to system to be operated automatically, while sufficing Definition 1., the definition of "efficiency."

From now on, call this number of bicycles needed at point X as Bb(X)

III. number of ports

We have defined in the previous subsection that the number of bicycles needed at point X, Bb(X)= Max[Mt1(X), Mt2(X), ..., Mtn(X)] by theorem 2. Now it is time to define the number of ports needed to also suite the 'efficiency.' Denote it Bp(X) It directly follows from the theorem 2.that

Theorem 3.(Corollary of Theorem 2.) *The number of ports that satisfies the definition of "efficiency," defined in Definition 1. equals to Max[-Mt1(X), -Mt2(X), ..., -Mtn(X)]*

Proof We see that by the construction, -Mti(X) = -Pti(X->1)-Pti(X->2)-... - Pti(X->n)-Pti(1->X)-...-Pti(n->X) = -|Pti(X->1)|-|Pti(X->2)|-... -|Pti(X->n)|+|Pti(1->X)|+...+|Pti(n->X)|. Hence this shifts the directional sign of Pti(X->1). Now in -Mti(X), the positive case represents there are more demands for parking the bicycles compared to the number of bicycles that are to be rented out. Hence, with the exact same reasoning that was used to prove Theorem 2., -Mti(X) represents the number of extra ports needed to suite people who are returning thier bicycles on time ti, assuming that originally, there weren't any extra ports other than already occupied ones in the station at point X. By taking maximum of these as we did in theorem 2., we get the minimum number of ports that could satisfy everyone who is trying to return his/her bicycle at every time step ti. Also, it exactly follows from the definition that the number if "efficient."

Together with the theorem 2, theorem 3 also defines the necessary environment for optimality ("efficiency")

V. More on Floating Population

We have talked a lot about the function P(x)i->j during the last few chapters. Hence, this function is tremendously significant in tracking the population and hence constructing the efficient graph model (V,E;T). However, before applying this function P(X) to many theorems and constructions of the model, we have not really discussed about the derivation of the P(X) value itself and validity of this concept. The first question that anyone would ask is how is this funtion P(x) derived despite the fact that we only have the data of people "passing by" specific point, just as what we have seen as an example in Figure 3 and Figure 4, but not the data of how many people actually move from a vertice to a vertice.

It is in fact we are only provided the data $\frac{d1}{dt}, \frac{d2}{dt}, ..., \frac{dn}{dt}$ for the vertices 1,...,n in set E of (V,E;T). Here, $\frac{di}{dt}$ means the number of people passing by vertice i during a single time step. Now, consider the very first time step t0 to t1, which is illustrated in the Figure 1..

We notice that in the very first time step, the number of float population at the starting vertice s equals to the summation of float populations of other vertices that are in the range of s. That is if s is a starting vertice and i,...,n are the vertices adjacent to s (in range of s), then $\frac{ds}{dt(0->1)} = \frac{d1}{dt(1->2)} + \frac{d2}{dt(1->2)} + .. + \frac{dn}{dt(1->2)}$ Then, it is simply from the definition of P(x) that $\frac{d1}{dt(1->2)} = P(s->1)t1->t2$, $\frac{d2}{dt(1->2)} = P(s->2)t1->t2, ...,$ $\frac{dn}{dt(1->2)} = P(s->n)t1->t2$ Also, this equation could be inductively extended, where we consider every vertice adjacent to the starting vertice as the starting vertice s′ in the next time step and so on. However, for the points i,j and a point x that is either in the range of both i and j or at least in the semi-range of both i and j, it is nearly impossible to figure out accurate P(i->x) and P(j->x) since the intersection gives computation too complicated.

However, from time step later than t2 and for the vertices that are not adjacent to the starting vertce s, it involves probability and more statistical approach to derive a rough guess of P(x).

I. Attraction rate and Degree

In previous section, we have defined the process of getting minimum number of bicycles and ports needed for each bicycle station. These are , again, denoted as Bb(X) and Bp(X). Now We have another definition.

Definition 4. *Attraction rate at vertice X, denoted* $B(X) = \frac{Bp(X)}{Bb(X)} \times k$ *for some weighted constant k*

The "attraction rate" here literally refers to the measure of "how attractive certain location X is compared to other places." In other words, it refers to the likelihood of people to choose to visit X leaving behind other choices, which are vertices other than X and a starting point. If we try to see the reasoning behind defining attraction this way, it is very clear. We know that if certain location is in need of a lot of ports, it means that a lot of people are riding bicycles to visit the place, and if it has need for large amount of bicycles, it then implies that more people are not tend to stay longer in that place. Hence, "attraction rate" corresponds to what we call ′degree′ of a vertice in usual graph notation.

However, this definition of attraction rate B(X) is under assumption that we know every value of the function P(x) between any pair of vertices in any given time step. Without limited information about the P(x) values, the best way we can go is to define the attraction B(x) as simply the weighted relative attractiveness in comparison to other points, which is
$B(x) = \frac{\int F(x)dt}{\sum \int (F(i)dt)}$ where F(X) is just a two dimensional graph of what we have in Figure 4. where y axis equals to the population and x axis equals to the time, for each vertice X, and the summation is in terms of different vertices i=1,...,x,...,n.

Now, the attraction rate B(X) enables us probabilistic approach toward the model. Under assumption that we have gotten P(x) values data and built the attraction rate B(X), we can derive the following formula.

Corollary (from the definition 4.) $P(i->j)_{tr->tr+1}$ = $\int_{tr}^{tr+1} F(i)dt \times B(j) / \sum_{i=1}^{i=n} B(i)$ (Here i=1,...,n are vertices that are adjacent to j)

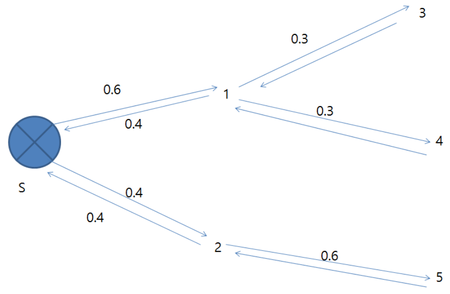
Here the only important thing to look at is that $B(j)/\sum_{i=1}^{i=n} B(i)$ works as a probability for a person at j, adjacent to i would choose i as a next destination.

In the corollary, we have only showed this way of defining P(x) between a pair of adjacent vertices in a single time step. However, in the next subsection, we introduce a way to measure P(x) between non adjacent overtices i and j.

II. Adjacency matrix and tree algorithm

One might notice that in the previous chapters, we only dealt with vertices that are in less than a semi-range distanced away from each other. In this subsection, we are going to see how tree matrix and algorithm enables us to see the interaction between vertices that are very far away. Suppose we have defined attraction for every vertices and all P(X) values that are assigned to each of edges in the system (V,E;T). Hence, we also have , for any pair of adjacent vertice (i,j), the relative attraction coefficient that we have used in previous theorem, which is $B(j)/\sum_{i=1}^{i=n} B(i)$. re-evaluate the constant k in order to have this coefficient to be less larger than or equal to 0 and less than or equal to 1. Let's call this the relative attraction rate coefficient and denote it Prob(i->j) Now for every directed edge connecting these vertices i and j, substitute P(i->j) with this coefficient. For example, we have as follows.

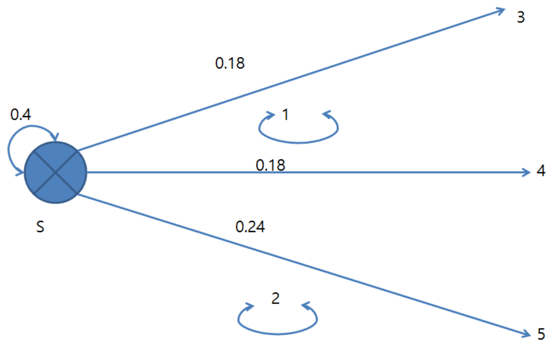
Figure 10: *subgraph of (V,E;T) with each edge substituted with relative attraction coefficient*



Suppose that in this figure, vertices 3,4, and 5 are not adjacent to each other, and vertices 1 and 2 are not adjacent to each other. Also note that values of edges coming out from the same vertex should add up to 1 by the construction of the formula $B(j)/\sum_{i=1}^{i=n} B(i)$. Now, by the independence of paths of the graph, a new abbreviated version of the graph with each edge describing the travel in two consecutive time steps could be created as follows. Note that this graph is obtained by multiplying every coefficient of the edge if one has to travel through that edge. We could check if it is valid by adding up every coefficient of vertices coming out from the same vertex, and see whether the sum is equal to 1. Hence, note that there are loops in this graph.

These loops in the above graph model represents the path leaving a point toward the adjacent vertex in time step 1 and coming back from the adjacent vertex to original location. Now, applying these steps inductively on the number of consecutive time steps, we get every Prob(x)for every possible travel routes during that time step, although it is in fact very true that there are more complicated loops formed, and these more complex travel routes make the computation slightly harder.

Figure 11: *subgraph of (V,E;T) with each edge representing a travel during two time steps*



Now, with the aid of another algorithmic process and the formula defined in the corollary of definition 4., we are now capable of computing every P(x) value for every pair of vertices (i,j), regardless of whether the points i and j are adjacent to each other or not. Let's put all the Prob(i->j) for every pair of vertices (i,j)on the n by n matrix where each columns and rows represent the the vertices of the graph. The matrix would look like follows.

Figure 12: *Adjacency matrix where each entry is equal to Prob(i->j)*

$$\begin{array}{ccccc} & 1 & 2 & \dots & n \\ 1 & 0 & \Pr ob(1->2) & & \Pr ob(1->n) \\ 2 & \Pr ob(2->1) & 0 & & \Pr ob(2->n) \\ \vdots & & & & \\ n & \Pr ob(n->1) & \Pr ob(n->2) & & 0 \end{array}$$

We have the diagonal symmetric matrix with each of diagonal entries equals to 0. Hence this is identical to how we define adjacency matrix of the directed graph (V,E;T) in usual graph theoretic notations. Also, note that this matrix can be transformed into a matrix with entries being P(x) values by multiplying a column vector of float average float population of first, second, ..., kth time steps. Depending on when is the first time that initial influx happens.

VI. More on efficiency and choice of candidate cities

Let's start with following seemingly obvious theorem

Theorem 4. *There exists the most efficient system (V,E;T) that satisfies the definition of efficiency in Definition 1.*

Proof If follows from Theorem 2. and Theorem 3. that we could obtain optimal sets V and E, and using these V and E, we could build up optimal model of full graph (V,E;T).

As we have seen, we have defined in Definition 1. efficiency is satisfied in terms of number of available bicycles and ports at each station. Hence, as stated in the theorem, we have proved that the definition is well-defined by finding a way to construct most efficient system of graph that suites the criterion in the definition. Now we suggest another potential definitions for efficiency.

Definition 5. *The model (V,E;T) is more "efficient" than the model (V′,E′;T) iff the bicycles in (V,E;T) are used for longer period of times than the bicycles in (V′,E′;T)*

This definition of "efficiency" is rather comparative. By this, I mean that this definition is useful to compare efficiency of two different models (V,E;T) and (V′,E′;T) but not to find the most efficient model because solely according to this definition, the most efficient such model would be a model with only one bicycle always in use for whole time step t0 to tn. Note that it is easier to think of the amount of time in use as the total cumulative distance that each bicycle travels.

Corollary *The model (V,E,T) is more "efficient" than the model (V′,E′;T) iff the bicycles in (V,E;T) move longer distances than do the bicycles in (V′,E′;T)*

Eventually, if we look at the travel of each individual bike, it is possible to deduce that bikes travel more if more vertices in its travel routes have positive M(x)(which we have defined prior to statement of theorem 2. and 3.) values at the time of the bike's arrival. This is because if the M(x) value is positive,then the additional bike that is being returned is likely to be used right away since positive M(x) implies the need for more bicycles to travel to another locations. Suppose that every bicycle move the distance d in a single time step. Then , the total distance that each bicycle travels is equal to k x d ,where k equals to the number of vertices that at the time of bike's arrival, has M(x) value larger than or equal to 0. Nevertheless, we arrive at the following conclusion

Observation *Being more "efficient(Definition 1.)" does not always imply that it is more "efficient(Definitin 5.)"*

This is very obvious. Suppose that there are two systems (V,E;T) and (V′,E′;T) where (V,E;T) suites more everyone in the city but (V,E;T) does not. By the construction, the only thing that differs is the number of bicycles or number of ports , or both, possibly. It is also clear by the definition that the model (V,E;T) has more number of bicycles or ports, or both. WLOG suppose that it has more bicycles. Then, if there is at least one vertex x where M(x) differs in some time steps ti and tj, then there are so called 'surplus' bicycles that are not in use during some r time steps. Hence, it moves r x d shorter distance than the bike which is in use during every single time step. However, if we let (V′,E′;T) be the model that contains the small number of bicycles, not enough to suite everyone, it could be the case that every bike in this model is fully used during every single time step.

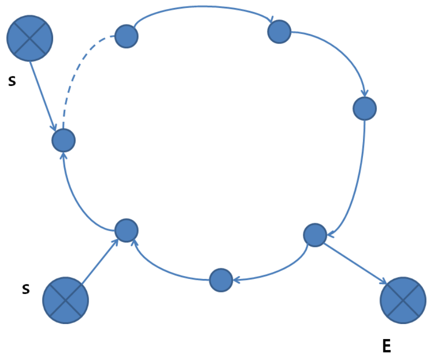
I. Macroscopics

Another significant feature is the macroscopic aspects of the candidate city for system installation.

We have seen ,in couple figures above,the rough picture of how the shape and structure of graph model would look like. This picture was initially based on the data and was readjusted to be placed in more systematic form that is more easier to be analysed. However, one important fact to note here is that our models that were depicted in the figures above were mainly based on the population data of weekdays in urban complexes. Now in this section, we will also try to have a rough guess of how it would look like if we are trying to install another bicycle system (V,E;T) that is operated during the weekends in complexes with tourist attractions.

Observation

Figure 13: *Outcome of simulation of algorithmic process that was used for Figure5 through Figure9 on Tourists complex*



Here, each of small circle represents the vertices E of graph. Also big circles S are strating points, while E are exit points. These starting points S behave similarly to the S in Figure 1 and 2, and ending points E are the points where people end their days in the city. Hence during the first time step, E contains no float population. We see that this outcome is more likely to be an euclidean trail. For that reason this version of the graph is not graph homomorphic to the one that was designed for urban complexes. Hence this graph cannot be transformed into the form of Figure 1 or Figure2.

Next shows a tool to see if the graph structure was properly adjusted to the geographic and demographic properties of certain cities. Horst Bunke, Peter J. Dickinson, Miro Kraetzl and Wlter D. Wallis introduce "clustering coefficients" and results of simulation based on it in their book "*A Graph-Theoretic Approach to Enterprise Network Dynamics*." The clustering coefficient here is the formula that values how tight is the local points adjacent around the vertex x of the graph is gathered around each other.

Here we define the *neighborhood* of a set of vertices consists of all the vertices adjacent to at least one member of the set, excluding the original members. Suppose we have graph generated by the neighborhood of x. Hence, denote numbers of vertices and edges of this graph as k(x) and e(x), respectively. The clustering coefficient is defined

$$\gamma(x) = \frac{2e(x)}{k(x)(k(x)-1)}$$

This coefficient equals the number of connections between neighbors of x divided by the maximum possible number of connections. As a result of that, it is great indication whether the adjacent vertices of vertex x are adjacent each other. In our model it decides whether the vertices that are in the range of a vertex x are in range or in semi-range of each other. Hence, it is very important factor that connects and expands the graph model.

VII. Conclusion and comments

It might have been the case that the problem of building an "efficient" model of public transportation system was unsolvable question from the beginning. The first reason behind this was being "efficient" or being "optimal" could mean a lot of things. Someone would interpret it as the equilibrium of strict supply and demand relationship in the market, while others define it as more of the choice problem and game between system provider and system users. However, what I have focused mainly on this thesis was the fact that this transportation system is rather 'public.' In other words, we are pretty much assuming that under certain amount of cost, we do not have to put that much effort into reducing the cost of building it. The more important part, I thought, was finding a way suite every or at least majority of participants' need by studying the way how potential users would behave inside the system both in macroscopic and microscopic perspectives.

I still admit that I stated a lot of assumptions in this thesis to try to prevent trivial geographical barriers from distorting the whole system. Also some might say that it is too hypothetical and non-applicable to real settings. Nevertheless, my goal was to find a way to set up a model with very limited information, and the simulations of this hypothetical systems printed highly accurate and valid results when run with the data set that I had. Please refer to Official census data on city of Seoul by Korean government(http://stat.seoul.go.kr/initinfo/) and archive data on float population of 1970's in city of Chicago and New York from U.S governmental census(www.census.gov/compendia/statab/).

The idea to apply graph theory in constructing optimal public transportation model for cities first came to me when I was first taking Combinatorics lecture in University of California at Berkeley. Hence, the prompt itself was inspired by 2010 High school Mathematical Modeling competition, during which I was gven a problem that asked me to analyse public transportation in the city of Chicago. Most importantly, I thank a lot to professor Chris Shannon(Mathematics, University of California at Berkeley) and professor David Aldous(Mathematics, University of California at Berkeley) for allowing me to visit their offices, present about the progress of research, and ask for additional help and any insights that enlightened me. Besides, all these great opportunities and great people who have inspired me, I hope this thesis could also help future undergraduate and graduate scholars who are willing to struggle with the optimization problem of public transportation or network system. In my own opinion, the beauty of this problem is definitely its "flexibility." By this term, I mean that this model could always be adjusted and applied to different problems such as optimal placement of Wi-fi netwroks, solution for traffic problems, understanding of the complicated structures of World Wide Web,etc. Finally, thanks all of my fellow readers for patience while following me in this thesis, and sorry for those who had descent amount of troubles trying to go through my clumsy definitions and constructions.

References

[Mark Koryagin, 2014] Game Theory Approach to optimizing of public transport traffic under conditions of travel mode choice by passengers *Transport Problem*, Volume 9 Issue 3

[Haray and Norman, 1953] Graph Theory as a Mathematical Model in social Science *University of Michigan Institute for social Research*

[Bunke, Dickinson, Kraetzl, and Wallis] A Graph-Theoretic Approach to Enterprise Network Dynamics. *Progress in Computer Science and Applied Logic*, Volume 24

[Wu and Chao] Spanning Trees and Optimization Problems *Discrete Mathematics and its Application*

[Clifford W. Marshall, 1971] Applied Graph Theory *Wiley-Interscience*

[Easley and Kleinberg, 2000] Networks Crowds and Markets,Reasoning about a Highly Connected World *Cambridge University Press*