Contents

Figure Contents

Table Contents

Executive summary

Bipartite network is the kind of network that originates from the field of pollination research, via the computation of some values based on the network between insects and plants, some results will be calculated for the protection of some species that might face the danger of extinction. From the angle of practical application, the limitation of such a model is evident. Once the dataset is constructed, results will be computed based on all the information from it. Namely, the dataset is always in a static state, which will also restrict the application coverage of this relationship model. Under such a circumstance, this dissertation aims at raising the coverage of the bipartite network. The original 2-dimension relationship network will be raised to a 3-dimension model, apart from group A and B like the relationship between insects and plants from the pollination network. A third axis, time, will be added into the network to make the model more compatible to the various field and datasets.

From the aspect of implementation, the principles of link variation of the matrix will be introduced in detail, including the appearance of new links and disappearance of old nodes. Besides, time evaluation is the innovation spot in the new model. Two different methods, feature list evaluation and Perez's curve evaluation will be introduced specifically. Finally, samples based on real data will be demonstrated for a better understanding of the construction of the model. Practical process such like data transformation, parameter calculation and threshold value configuration will be introduced based on the whole sample. After the data construction, a result analysis will be provided to have an objective evaluation of the construction of the new model, which including the criterion from different aspects such like sequence and life span. In addition, improvements for the new model will be given, some of them are based on the result of the real data analysis, and others are listed for the better performance. Apart from these, some potential applications out of data analysis are also introduced to show the wide coverage of the new model. On the other hand, since all the improvements are on the theory level, implementations are necessary for evaluation.

Finally, a short list of bullet points that summarise the main contributions and achievements in the project will be listed as follow:

- The basic concept of the bipartite network is expanded with a new variation, 3-Dimension bipartite network.
- Two different methods for the time evaluation in the new model are created, and their performance are analysed.
- Introduction of real data processing is given for a better use in other fields.
- Improvements for the model quality are given for future enhancement.
- Cross-field applications are introduced to provide some inspiration in the area apart from data analysis, which also expand the coverage of the new model.

Chapter 1
Aims and objectives

Network relationship is a kind of concrete representation method that can demonstrate quite a lot of organizations and communities from area to area which is close to our daily life. Since a network relationship can be defined based on the specific demand, therefore, huge amounts of variations exists to solve different sorts of problems from macro levels to micro levels. Generally, a relationship network is composed of entity nodes and edges, thus, such a description can be easily transformed to the form of graphs or matrixes which is not difficult for understanding. In other words, implement some analysis based on the network relationship has the advantage of operability and measurability.

Relationship networks can be distinguished according to the difference of network structure such like unipartite network, bipartite network, neural network and so on. In this piece of work, we focus on the research of bipartite network, and aim at some further exploration based on the current theories and research results. Specifically, the work will upgrade the bipartite network from the static level to a dynamic level, expanding its function and application field. From an overall view, the researching related objectives of the work can be listed from the following aspects, which can be considered as the main line of the research.

1.1 Construction of a dynamic network model

As an expanding version of the existing bipartite network, it is of great importance to have a good understanding of its original concepts, including some measurement parameters and their related calculating algorithms. Although the aim of the work is to create a new network model, the existing materials cannot be neglected, they are the foundation of the whole work. Besides, since the work is closely related with something which has not been mentioned before, therefore, in order to construct a complete network model, some points are necessary to be specifically introduced such like the new added elements, their affect on the new model, implementation methods and parameter configurations. Finally, a basic version will be given for the sample analysis and evaluation.

1.2 Real data simulation

Practical application is the final goal of a model. Therefore, in order to reach this goal, several steps should be executed for both accomplishments and improvements. First of all, an objective analysis of the dataset used in the original bipartite network should be given, which would be a useful reference to define the dataset for the new model. Secondly, since the model is constructed for application expanding, different dataset

with various features should be taken into consideration and being deeply analysed. Such a practice is necessary for the polymorphism of the model, namely, it is more ideal to construct a model which can deal with different situations. For the last, the source of the datasets for the sample analysis is significant, only build a convincing data matrix can guarantee an objective evaluation of the new model.

1.3 Results evaluation

This is a core section for the work, as a new constructed model, it is reasonable that it might perform well or terrible, every result is possible to happen, most important of all, a detailed and actual evaluation should be prepared for the further research or improvements for the model in the future. Specifically, practice from several aspects could help to enhance the quality of the result evaluation. First of all, results obtained from the model can be compared with the actual data, which can be regarded as the most intuitive material for evaluation. In addition, results can be analysed with the help of some existing theories. For example, if some theory had been proved that there exists a special relationship between two targets, therefore, the more close the result from the model to the theory, the more convincing the model could be. Furthermore, testing the model with different parameters is a reasonable practice, from one hand, iterative testing with changing values helps to obtain better parameters for the model, from the other, the series of testing assist the evaluation of the work to be more thorough. It can reveal the difference between different value settings and their related reasons that lead to such a phenomenon.

Apart from what has been mentioned above, evaluation of another aspect which is also indispensable that should be demonstrated. Future improvements and application fields description is also an attractive section of the work, the former information will provide some potential ideas as the reference for further model enhancement, while the latter will define some criterions which help the users to evaluate if the new model would help to solve their problem. In other words, if the model is designed with a higher compatibility, the wider area it could be applied into, which is a more ideal result that being preferred.

With all the aims and objectives displayed above, it is necessary to introduce the techniques and materials available at the moment, which would help to have a better understanding of the latter chapters. Techniques and materials are mainly from the following angles:

1.3.1 Network aspects

Current theories and algorithms from the network area is the foundation of the new model. It is not scientific to build a model about the network without any former research results from this area. On a more objective sight, the use of existing research findings about network, or especially bipartite network is reasonable. It means that some classical ideas and algorithms will be applied and recombined as the basis of the new model.

1.3.2 Upgrading aspects

This is a huge information resource pool because so many potential theories can be applied to the new model for upgrading. To raise a simple instance, theories from the areas like artificial intelligence, business economics or even psychology can be added into the new model if needed. There is no strict limitation in this aspect, as long as the knowledge seems to be useful to the model construction. From the view of real practice, although it is impossible to implement all the potential theories to the new model which might be hampered by the implementation difficulty, the good ideas will be shown on the improvement section for reminding.

1.3.3 Programming aspects

Since the model is implemented via the form of program, hence, improvements might also be explored from this aspect. Some algorithms or data structures might enhance the efficiency of data processing. Improvements of this kind should be applied based on the program features of the model such as data scale or algorithm in use. On the other hand, during the use expanding of the model, some programming technique might be applied, such as network programming or multithread programming for example. Overall, improvements from this aspect are influenced by the programme development procedure and data features as well.

Based on the introduction of aims and objects above, the position of the work can be given and clearly displayed through the form of a figure:

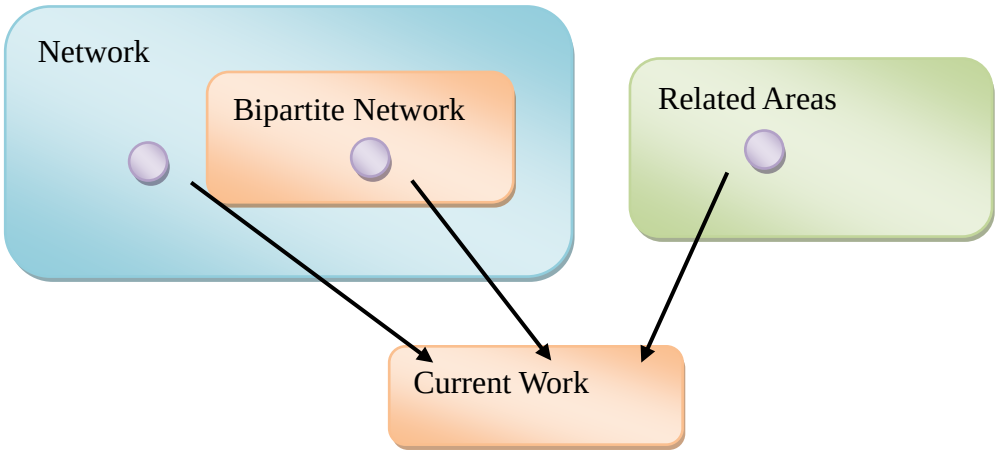


Figure 1. The knowledge composition of the work, the research result of this work will be sorted into the group of bipartite network. While during the processing of the research, not only the existing background from the bipartite network, knowledge from the general network and some other subject will be applied into the work.

With the general idea about the construction of the work, the application aim will be briefly introduced here. As is known so far, bipartite network is a kind of static network that can reveal some potential information from some of the characteristic

value. A typical example could be the pollination network, with a piece of data available, some value could be calculated to show that if it is urgent to protect some of the species from either plants or insects. By comparison, dynamic bipartite network is quite different from the static version, the results given from it will be based on the simulation of the actual development of the dataset of the selected area. In other words, dynamic bipartite network is trying to predict some potential results with the information which has not happened yet. Generally speaking, if the traditional bipartite network can be regarded as a 2-Dimension network, the dynamic bipartite network is a kind of 3-Dimension network, whose third axis is time.

Based on such an orientation, the application of the dynamic bipartite network could be wide. Specifically, the objective can be separated into the following parts, first of all, a detailed data analysis will be given in order to provide a better understanding of the feature of dataset that can be applied into the new network. Besides, some real examples will be demonstrated to show the application in other fields apart from pollination network. Finally, a general growing route will be introduced to describe the probable future of the network theory which aims at offering a comprehensive concept to the users. On the whole, from the view of actual application, the aim of the work focuses on providing an easy comprehension network model for more users from all the aspects that they are interested in, such like data feature, sample execution, result analysis and development evaluation.

Chapter 2
Background and context

In this section, the background of the network field will be introduced from the basic concepts to some typical parameter values, which would help to the understanding of the ideas and principles of this work. Since the dynamic bipartite network is a new variation from the network area, therefore, backgrounds and contexts in this section is closely related to it.

2.1 Basic concepts of bipartite network

Bipartite network is a special kind of network that displays the relationship only between two main sets, such as teachers and students, users and application softwares, etc. Besides, bipartite network holds the basic concept same as the unipartite network like vertex, nodes, direction, weight, and so on. In this part, the concepts of the bipartite network will be given from the basic points to the complicated notions.

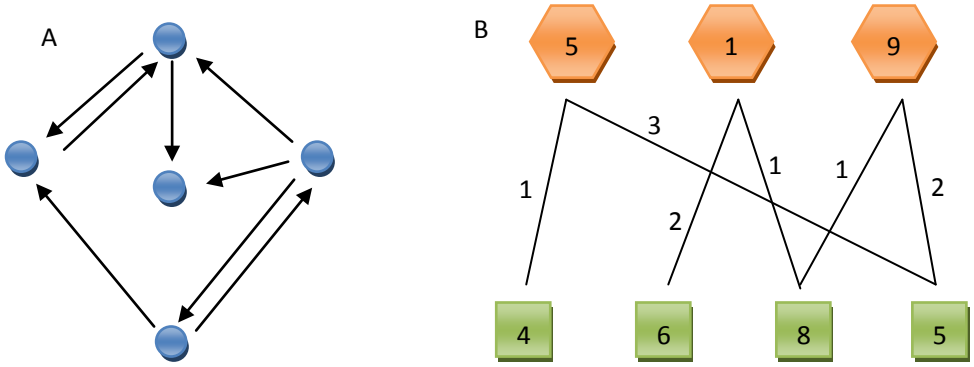
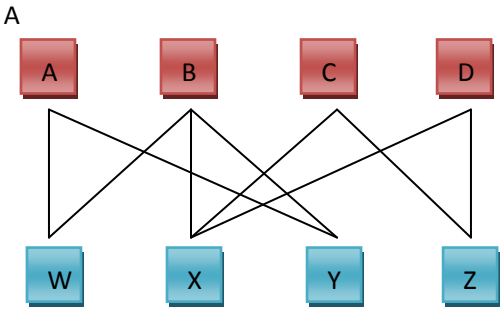


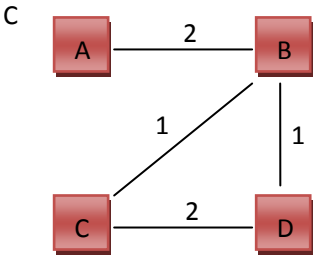
Figure 2. The basic concepts of unipartite network and bipartite network. (A) Unipartite network, the dots are defined as vertexes or nodes, where the lines are defined as edges, this network is a directed network since there are arrow exists, namely, node A can point to node B and vise versa, they are not same with each other. (B) Bipartite network, it is easy to identify that hexagons and squares are two different sets, they do not connect with each other. The network is not a directed one because there are no arrows in it. Most important of all, vertexes and edges are labeled with numbers, such a kind of network is defined weighted network, numbers of the vertexes and edges represent the scale of vertexes and amount of links separately.

Typically, one of the useful transformations of the bipartite network is projection [1], because during the network analysis, information from the relationship between the nodes from the same group is also important for the prediction.



B

	A	B	C	D
W	1	1	0	0
X	0	1	1	1
Y	1	1	0	0
Z	0	0	1	1



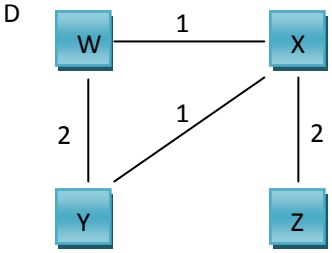


Figure 3. The basic transformation of bipartite network. (A) The original topological figure of a bipartite network. (B) The matrix of the link of the network, which can be used to calculate the nestedness of the network. (C/D) The projection from bipartite network to a weighted unipartite network of the red/blue set. It is more evident to have a clear understanding of the relationship of the nodes from the same set, which might be useful for the further work like near-future prediction.

Apart from the basic concepts mentioned above, the complex notions from the former works about bipartite network which are also useful to the construction of the new model will be mentioned as follow:

2.1.1 Modularity

This is a parameter to evaluate the density of clusters in a network, in other words, in a network with a high modularity, intensive links can be seen between the nodes in the same cluster, on comparison, there are less links between the nodes from different clusters.

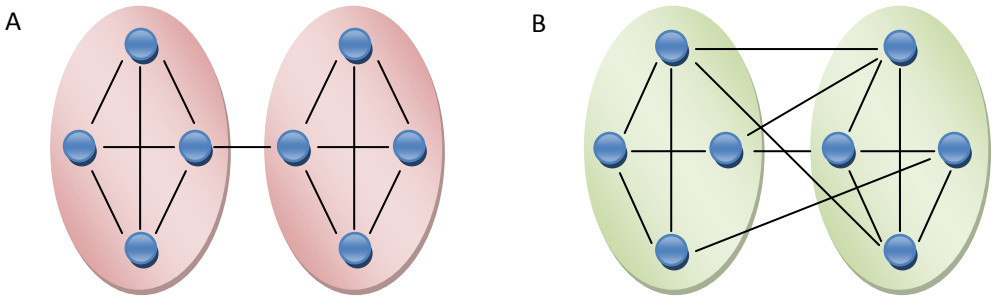


Figure 4. Concept of modularity. (A) A network with high modularity, two clusters can be easily found. (B) A network with low modularity.

2.1.2 Robustness [2,3]

This is a parameter to evaluate the ability of survival if a certain node in a network is removed. Specifically, survival here means that in a network, node A can still find a path to node B even if node C in the whole network is deleted. Theoretically, if the robustness of a network is high or strong, it is hard to be divided into several sub-networks if some of the nodes are deleted by chance. Take a practical example, if a computer network system has a strong robustness, even if one of the computers is being attacked, the rest of the system is still working, but if it is weak, the system will be broken instantly.

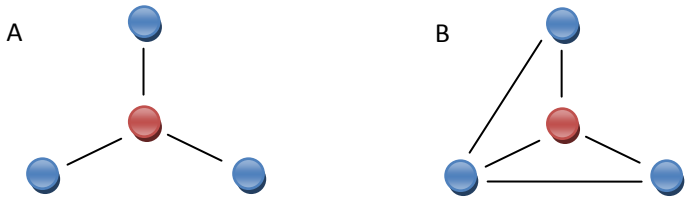


Figure 5. Explanation of robustness. (A) The network will lose all the links of the blue nodes if the red node in the middle is removed by chance, thus, its robustness is comparatively low. (B) In this network, even if the red node is removed, blue nodes are still able to connect each other through different paths, namely, it has a higher robustness.

There are some parameters that can imply the robustness of a network such as degree centrality, closeness and between-ness. Degree centrality depends on the number of links from the specific node to its neighbours. Closeness is the value that indicates the average distance from the selected node to all the other nodes in the network. Between-ness describes the frequency that been passed from all shortest path between each pair of nodes. In the unipartite network, such of these measures would help to evaluate if the lost of a specific node will lead to the failure of the network. While in a bipartite network, such a character can be measured with a new parameter, nestedness, which will be introduced in the following part.

2.1.3 Nestedness [4,5]

This is a concept from the field of ecologic, originally, nestedness is a measure representing the constructions of species of the continental biotas, which lives in different separate habitats such as islands and landscape fragments. Generally, in a normal nested network, the sort of species of the small assemblages is strictly an ordered subset of the large assemblages. In other words, similar to the concept that the closer the island to the mainland, the more species of animals it will hold. If a network has a high nestedness, it means that we can firstly find a node from a group that connects with a certain amount of nodes of the other group, and one node can be found which connects the nodes of the subset of the former amount after another. The higher the nestedness is, the rule will be displayed more strictly. More formally, in a particle network, nestedness is a measure that demonstrates the likelihood of a link from the low order node to a comparatively higher one.

	A	B	C	D
W	1	1	1	1
X	1	1	1	0
Y	1	1	1	0
Z	1	0	0	0

	A	B	C	D
W	1	1	1	1
X	1	0	1	0
Y	0	1	0	0
Z	1	0	0	0

Table 1. The two separate tables display the difference between different nestedness. Assume that A-D represent a group of different islands from close to far to the mainland, and W-Z represent 4 kinds of animals. If animal W lives on the island A, 1 is labeled on the corresponding position of the matrix. Since island A is the closest to the mainland and D is the farthest, a higher nestedness means that every animal on the farther island has a higher property to be seen on the closer island, obviously, according to the tables above, it is easy to draw the conclusion that the table on the left has a higher nestedness than the right one.

- Different measurements from Ulrich and Atmar [4]

There exist some popular ways of the measurement of nestedness, such like “discrepancy” and “matrix temperature”, the former one is easier to compute while the latter one is comparatively more popular but needs more solid mathematical foundation.

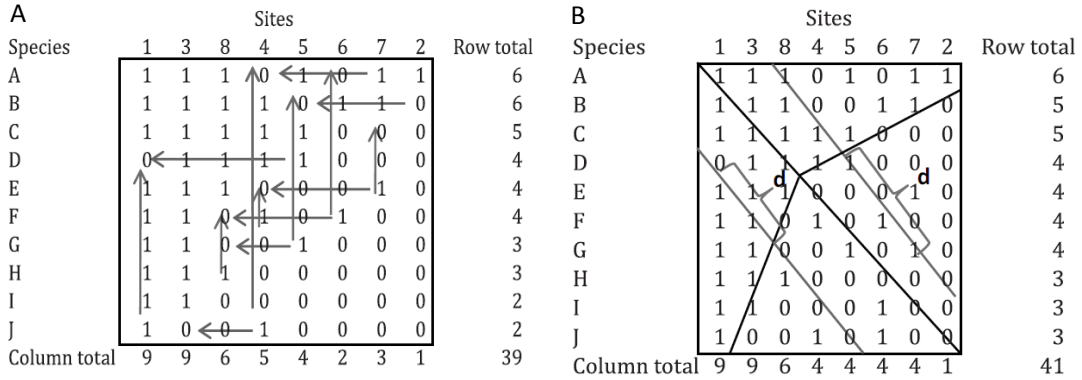


Figure 6. The brief explanation about the recommended nestedness measurements. (A) Discrepancy computes the minimum numbers of shifts from the current state to the perfect nested matrix, in the given case above, the minimum value of shifting is 7. (B) Matrix temperature is the normalized sum of squared distances d of holes (0s) and outliers (1s) from the isoclines along the matrix diagonals. Figure 6 is cited from [4].

- Lee’s suggestion about computing nestedness [6]

Although the matrix temperature is quite popular, it is not easy to calculate since it imports the concepts from linear algebra. In Lee’s paper, a more effective method to measure the nestedness is introduced, which is considered having as good as or even better performance than the matrix temperature. Briefly, the computation of nestedness is defined as follow:

$$N = \frac{1}{k(k-1)} \sum_{i=1}^{k} \sum_{j=1}^{k} \frac{\sum_{l=1}^{k} a_{il} a_{jl}}{min(d_i, d_j)}$$

where “di” is the order of node i, $\sum_{l=1}^{k} a_{il} a_{jl}$ is the sum of the common neighbors of certain node pair i and j, k is the number of nodes of the whole network. Aiming at the project, it seems to be useful to handle the projection of the bipartite network since such a method is suitable for the unipartite conditions. It could help to deal with the chain reaction with the appearance of a specific node.

- Rodrıguez-Girones’s new algorithm [7]

The appearance of another algorithm to compute the nestedness of the network tends to raise some other drawbacks of nestedness temperature especially aiming at the fact that the calculation of the “temperature” is demanding for the reorder of rows or columns and some random values, which could influence the accuracy of the final result. In Rodrıguez-Girones’s paper, a new algorithm BINMATNEST is introduced to order a better solution of the problem. Generally, this measurement is based on the genetic algorithm to decide the reorder of rows and columns, in order to create a minimum value of the matrix temperature. Hence, such a method is a great choice for the refinement of the analysis results.

Based on the knowledge introduced above, a main concept of bipartite network and its related functions and values have been established. Concepts in this section will be partly applied into the construction of new model and the potential future improvements as well in the following parts.

Chapter 3
Description of the work

In this section, the description of the whole work will be specifically introduced by sequence, from the general idea of construction, the major theories or algorithms in use, to every detail that cannot be neglected.

Focusing on the application of the pollination network [8], the majority of its purpose is to find the potential extinction species via computing some special values such like nestedness. With respect to dynamic bipartite network, it is based on the bipartite network theories. On the other hand, its own features are evident as well. Generally, the basic model can be demonstrated with the following structure figure, and the details will be given based on the sequence of the figure:

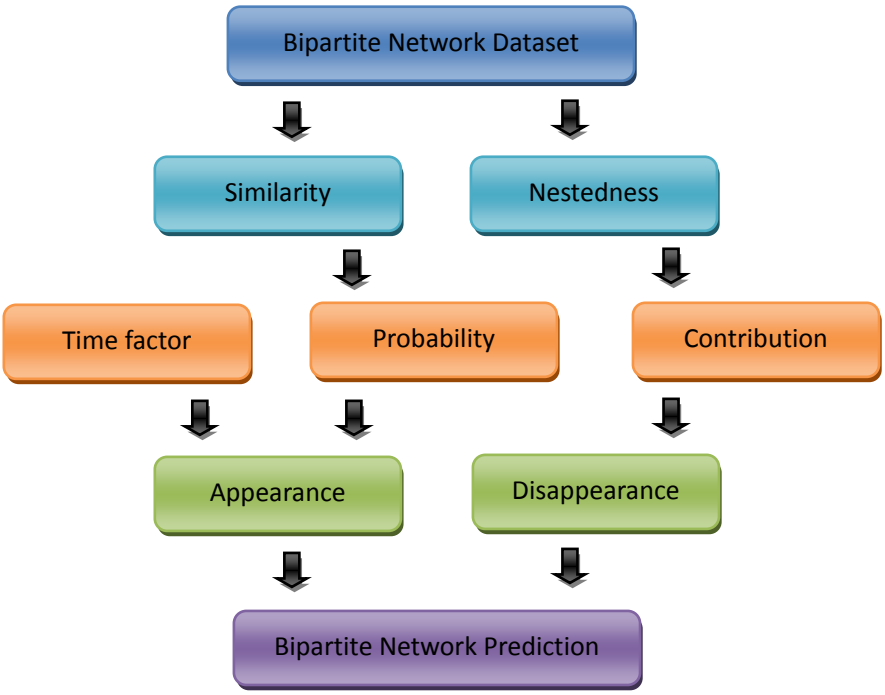


Figure 7. The main structure of the basic dynamic bipartite network model from the source dataset to the methods to obtain the results.

3.1 Network Dataset

Since the bipartite network model was originally used in the pollination network, it is necessary to have a comprehensive analysis of the features of the data from the ecosystem, in order to have a better assessment of the data from the other areas. Hence, in this part, a detailed analysis will be given to reveal the difference between

datasets of different areas.

Before exploring the data from other fields, it is of great importance to recap the features of the dataset from pollination network specifically. The features will be labeled with a phrase and be described with a further explanation as follow:

- Two main sets: Plants and insects are the only two main groups in this bipartite network, which is also a key rule for any models in other areas.
- Number of nodes: In a specific pollination network, the sorts of insects and plants are kept in a certain amount level. Besides, the quantity of the sorts is up to around one hundred since the chosen species have a bottom threshold.
- Mutualistic: In the pollination network, plants and insects rely on each other to live, in other words, from the view of computer science, the life span of the vertexes of one group in the network depends on the interaction with the other group, namely, if a node extincts in the bipartite network, it implies that the weight of its link with others is below a certain level.
- Choice stability: This feature depends on the characteristic of the ecological field, as long as the insects and plants live in an ideal environment, comparatively, insects will not change their links to the plants within a wide range with a high probability. Except for the situation such as a forest fire that will destroy the environment which would make some species disappear.
- Long life span: As a part of the whole ecological system, the life span of the pollination network of is very long, the species keep their generations with the loop of the seasons, similar to the idea of “Choice stability”, the network will be well kept if the objective environment is suitable for the species.

With the brief analysis of the insects’ world, we can focus our sights to the world we are more familiar with. The common dataset we can see is quite different with each other. For instance, if we pick up the friend relationship from Facebook, we can find that it is a huge unipartite network with millions of nodes and directed edges. Besides, the updating frequency of it is far faster than our imagination. For another, we put the attention on the situation of the relationship between customers and products. It is obvious that the model can be constructed as a bipartite network, but still, there exist some features are quite different with that of the pollination network. The core reason for the difference is us, the human beings. People are always interested in trying something new, besides, price will influence the purchase tendency, such as the discounts.

A nice example will be given here just for a reference, since it follows the features mentioned above. We choose the relationship between authors and the papers in the field of computer science, furthermore, the subject of papers can be classified to the sub-fields such as network, cryptography, database, graphics, etc, or they can be classified more specifically such as bipartite network. Two groups, comparatively small scale according to the specific need, classification of the nodes, and avoid the

unstable human factor (although author is one group of the two, but the professionals seldom tries to change their research field because it is not easy to make some research attractive if one is not familiar with that knowledge). Focus on this spot, a reasonable bipartite network can be built to research the phenomenon like technique development and extinction according to the potential information from the bipartite network.

Apart from the application itself, another point which is necessary to be mentioned from some existing paper is the source of the dataset of this project. Have an overall view at a specific example about economy [9], GDP and trading statistics of different countries are needed, although from the angle of an ordinary person, figures of this kind are tedious and a little bit mysterious, actually they are not difficult to obtain, they can be found from the website of International Monetary Fund (IMF) and International Merchandise Trade Statistics separately. Figures from such a kind of data source are accurate and credible, which correspond to the requirements of the dataset. Especially in the model of dynamic bipartite network, since we need to predict some future results from the dataset, in order to evaluate the accuracy of the prediction, objective past data are necessary for the comparison between program outputs and widely accepted figures.

As for the description method of the dataset, the most ideal way to demonstrate the relationship between two groups is using a matrix, such a practice has several advantages from different aspects. Obviously, it provides a direct two dimension table which is clearer than a link figure especially the amount of nodes and edges are huge. Besides, with the help of the matrix, it is easy to transfer the matrix to a graph when needed. Most important of all, this is a preferable way for the program to deal with the dataset and calculate some values.

3.2 Nestedness and contribution [10]

In a network, it is intuitive to imagine that the removal of a strong contributor tends to decrease overall network persistence more easily than the removal of a weak contributor. However, strong contributors to collective persistence do not gain individual survival benefits but are in fact the nodes most vulnerable to extinction. Obviously, this is a new paradox into the study of the persistence of cooperative networks because it does not follow our subjective judgment. The author proved this phenomenon by analysing a 15-year time series of the interactions between designer and contractor firms in the New York City garment industry. As with the ecological networks, a firm’s survival probability decreases as its individual nestedness contribution increases.

In order to judge a persistence of a specific node in the network, the definition of contribution of the node should be clearly demonstrated. The measure quantifies the degree to which the overall nestedness of the network compares with the value obtained when randomizing just the interactions of that particular node.

Mathematically, this is defined as:

$$C_i = (N - N_i^*) / \sigma_{N_i^*}$$

where N is the observed nestedness of the network, N_i^* and $\sigma_{N_i^*}$ are the average and standard deviation of nestedness across an ensemble of random replicates within which the interactions of node i have been randomized.

With the help of the picture, the way to quantify the measure can be explain more directly:

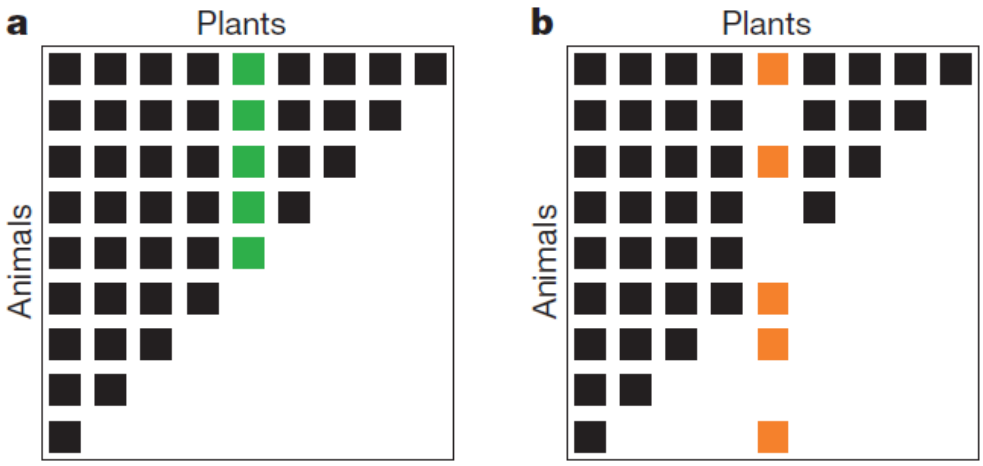


Figure 8. Explanation of contribution in a bipartite network. (a) The original network, the column in green is the node of plant we are interested in. (b) The comparison group of network to help the computation of contribution. Figure 8 is cited from [10].

When we get a dataset of the link between two groups, it is not difficult to display it with a matrix and get its nestedness, just like Figure 8(a) above. If we want to measure the contribution of a specific node, we should remove all the links of the node and replace the same amount of links to all the possible nodes randomly, like Figure 8(b). With a new network pattern, another nestedness can be measured. After the repeated experiment, we can use the set of nestedness value to calculate the average and standard deviation, and thus compute the contribution of the node i which is exact that we are focusing on. During the processing of computing the contribution value, the only thing we should take care of it the number of the time of the experiment, intuitively, it depends on the scale of the matrix.

Overall, the more a node contributes to nestedness, the more likely it is that its loss is detrimental to the network's persistence. Nestedness contribution therefore represents a key measure of the degree to which each node's interactions work for or against the long-term persistence of species in the mutualistic network. Specifically, in the practical application in other fields, we compare each node's nestedness contribution to its survival probability, in general, the more a node contributes to the architecture

of its network, the greater its probability of extinction.

3.3 Similarity and probability [1]

Apart from the macro measures of a network, micro measures also exist, apparently, it is not about the view of the whole network, instead, it concentrates on the potential behavior of one specific node in the certain group. What we display here is the idea of personal recommendation. In a bipartite network, assume that node i of group A has built links to several nodes in group B, but how about the probability that node i will connect to the node k in group B that has never been connected before? That is what personal recommendation could help.

Here we use one of the most popular method to solve this problem, namely, we use the concept of collaborative filtering, which is based on the result of similarity between the nodes from the same group. The formula of the similarity between two nodes is defined as follow:

$$S_{i,j} = \frac{\sum_{1}^{n} a_{ik} \cdot a_{jk}}{min(d_i, d_j)}$$

where d is the degree of the node, a stands for the linking status between the nodes, if i connects with k, then a of i and k is one, otherwise, it is zero, n is the number of nodes of the group that does not contain node i and j.

With the definition of similarity, now we can define the measure of the probability that if node i from group A is interested in the node j of group B, here we define P as:

$$P_{i,j} = \frac{\sum_{l=1, l \neq i}^{m} s_{il} \cdot a_{jl}}{\sum_{l=1, l \neq i}^{m} s_{il}}$$

where m is the number of nodes of group A. To explain this formula more clearly, we firstly calculate the sum of the product between all the other nodes in group A apart from node i and all the similarity of every node in group A and node i. The sum is divided by the sum of the similarity between nodes in group A. According to this formula, we can make some conclusions that P can be influenced by either the degree of node j or the similarity between i and the other nodes. As a result, the bigger of the value of P, the higher probability node i has to build a link between i and j.

3.4 Appearance and disappearance

With the data represented in a matrix, appearance and disappearance of the links between nodes from two groups can be influenced by probability and contribution separately, which enable the matrix to be a dynamic state. In reality, since the edge between nodes needs certain time to be constructed, therefore, such a characteristic should be reflected from the linking status matrix. In order to satisfy such a demand,

two main methods dealing with time will be introduced, specifically, the concept of feature list and Perez's curve [11] can be applied into the network model.

With respect to the feature list, there are two main concepts that will influence the prediction results of the model, namely, the selected feature items and their related values. From the angle of one group, taking computer science technology as an example, intuitively parameter type varies between different techniques. For instance, cipher text's length is an important value to cryptography, while 3D visual effect algorithm is significant to a 2.5D technology. Evidently, such kind of information is useless to a prediction because they are either common or spreading-related.

On a general review of the dynamic dataset, parameters like cost, efficiency, necessity, developing time or usage difficulty will influence the spreading speed of a specific released technology, which will result in its life span. Therefore, parameter selecting should be pertinence related to the certain field. Besides, value setting is also a key section, basically, there are two main methods for value setting. Firstly, level definition can be a typical practice, level definition like "low", "medium" or "high" can be used for parameter evaluation, and each level are related to the specific value which helps for the calculating of prediction. This is easier to operate with a low cost, but on the inferior side, the value setting is comparatively subjective and the value is not that accurate that can reveal the differences between nodes from the same group. For another, a more objective but costly way is to get the detail value for computing, such like time consumption for a specific operation or selling price for the entity with the certain technology. Since some of the parameter like technique commodity is hard to be set, thus it is more reasonable and objective to use a hybrid solution.

Apart from feature list, some other method is also alternative. In reality, it is almost impossible to find a situation that a certain rising is following the form of linearity. Since the dataset is chosen from the areas apart from ecological system, thus even the predictions should follow the features of the specific field, otherwise, they would be considered as useless. In order to make the results easier to translate, hence we import some concepts from the area of technological revolutions to upgrade the timing determination.

For different kinds of technological revolution, it is obvious that their life span are quite different. According to figure 9, the installation period and development period both cost about 20-30 years, intuitively, this is the description for some traditional industry like metal or energy. New technology like computer science is developing far faster than them, especially in the recent years. However, the shape of this curve is still significant to the evaluation of the prediction results because the development tendency of every new produced technology is very similar, in other words, people will accept the new comers following the steps of Perez's curve, the only difference is the developing speed. Recently, old technologies are easier being substituted by the new ones, the reason for this phenomenon is evident, the information is spreading too

fast, once some surprising technology is released, it will be spread to every corner of the earth overnight, after that, the relevant ideas, no matter from its supporters of protesters, will be published as quickly as we imagine, hence, the life span of the recent techniques are in a shorter cycle.

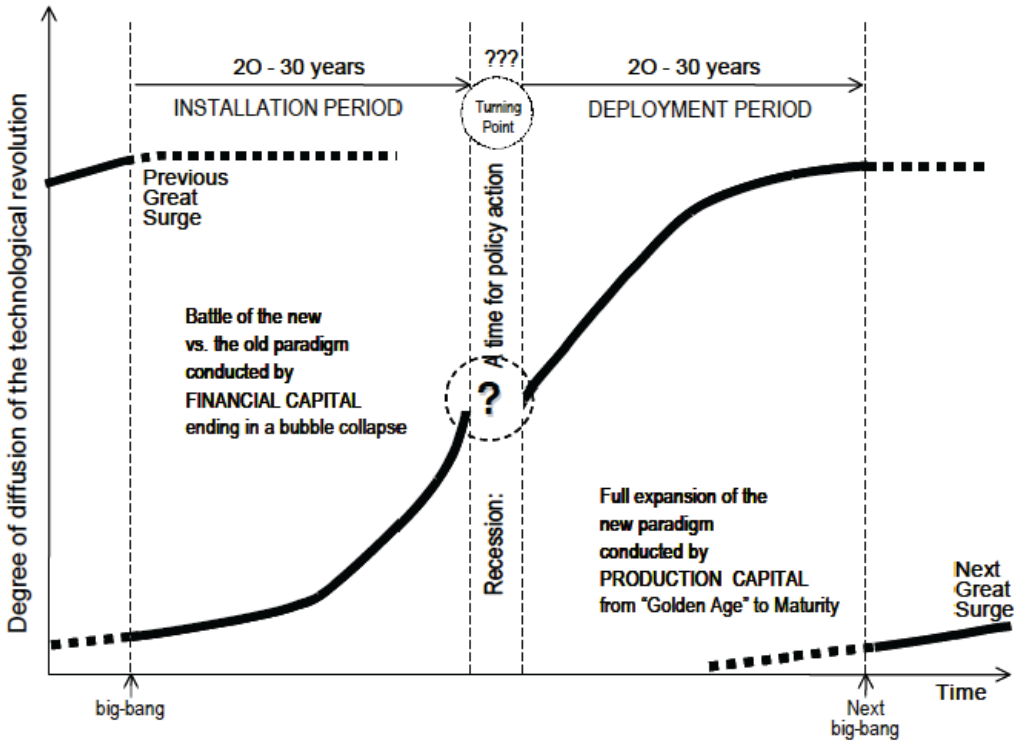


Figure 9. Carlota Perez's description of technological revolutions. Mainly the curve rises fast in the first half of the period, after meeting the turning point in the middle, it begins to grow in a slow speed. Figure 9 is cited from [11].

It is not easy to set the parameter of time span in the dynamic network model, but via focusing on the curve of the graph, there is some relationship between the link numbers from the network and the shape of the curve line, which will reveal some information about time.

Maybe it is a little bit difficult to give an accurate answer of the life span of a certain technique, however, it is comparatively more convenient to give a line and spot the location of the development according to the growing speed. In other words, if the numbers of the edges grows slowly after the appearance of the node, that means the technique is not widely accepted; if it grows fast between periods, it implies that it is being popular at the moment; when the growing speed becomes slow again, generally, the number of links is becoming stable at a high level, according to Perez's theory, some new technology is probably on the way to a new revolution.

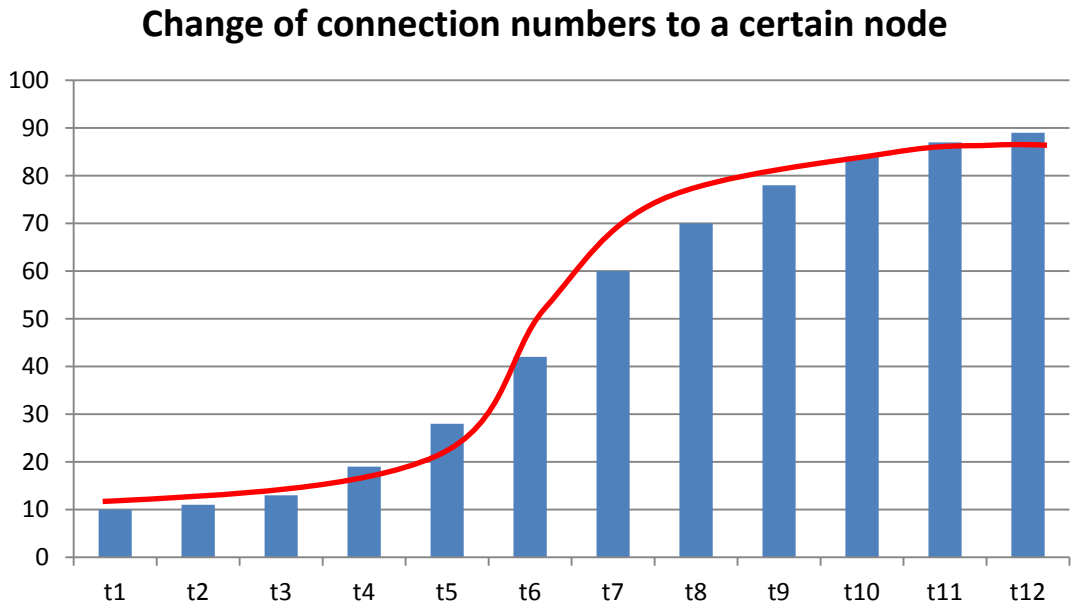


Figure 10. The figure of the number of the links to a specific node. Ideally, with the time passes by, the number of edges collected in different time can be shown with a similar curve from Perez (red line in the figure).

On the whole, the application from Perez's theory is useful in the building the model, specifically, such a practice translates some abstract concepts to something specific and can be computed this is an important point for a model construction.

Disappearance is controlled by the contribution of a node, unlike the appearance of the links, contribution is unrelated with other facts. However, since eliminated nodes will be substituted by new comers, thus, configuration for new nodes should be defined for the continuous execution of the dynamic network model.

3.5 New coming node requirements

According to the law of technology development, when one technology is widely accepted and increasing slowly on application amount, it implies that some new technology will appear and begin to replace the existing one. With respect to the angle of the matrix, when one node has big amount of links with the nodes from the other group, it will be substituted by a new node with a high probability. In order to simulate the matrix development more smoothly and objectively, the initialisation of the new node is of great significance. Based on the structure of the data, two main operations are necessary for the initialisation. From one hand, the feature values should be given by the node manufacturer since no one else can tell the relative features. From the other, the original links with the nodes from the other group should be given, otherwise, it is impossible for the matrix to spread the links with an empty initialisation, and it seems to be inconsistent with the facts. As to the weighted matrix, it is more difficult to give the original figures because the 0s and 1s should be replaced by specific numbers, which should be roughly evaluated.

3.6 Time Factor

Time factor is a significant component to upgrade the bipartite network from the static level to a dynamic level. Thus, in this part, the various concepts about time and their related measure methods will be introduced. Overall, time factor can be defined in different levels, which will be briefly demonstrated as follow, and the practical application depends on the features of the dataset.

- Short time period: Roughly, time span from minutes to days can be defined as a short time period. It is suitable for the datasets which have a high updating frequency, such as the number of some posts shared by the users, if we check a certain post a week after it was posted, it might already be forgotten by the users.
- Long time period: Roughly, time span from weeks to years can be considered as a long time period. Scenarios like pollination network, the edge between the certain insect and plant may not change even by years. Hence, as long as nothing special happens, there is no need to monitor the data so frequently.
- Life span: This is a parameter that depends on the dataset we choose, we define life span as the time from arrive to extinct of a new comer, thus, the life span varies from field to field, in the ecological environment, a specific species of plant can keep alive for hundreds of years under ideal conditions, while in the field of operation system, for instance, even the one of the most popular operation system Window XP, Microsoft will stop the support of it in 2014, namely, its life span is just less than 20 years. In this project, life span is just a reference for the controls of data processing, more care will be take care of during the special periods.
- Processing span: This is another parameter that will rely on the features of the selected dataset. Processing span is the time that the system consumes from collecting the data to making a conclusion, intuitively, processing span is longer than the time period and shorter than the life span, otherwise, the conclusion will be not convinced enough or out of date.

3.6.1 Timing representation

From the angle of model design, the results are originally displayed in the form of steps, therefore, it should be translated into a specific prediction which is related with actual time. Based on the design of the original version of the dynamic network model, since we get a result step by step from the model, it means that the sequence of matrix development is simulated by the model, thus, when the time consumption of one step is confirms, the whole prediction can be constructed. For instance, if the result contains the information like “step 1: A will happen” and “step 4: B will happen”, suppose that A happens three months later from now on, B will probably happens after another nine months. There is no doubt that the time span varies between predicting objects, sometimes the difference can be huge enough like days and decades, but since the developing sequence is comparatively stable, thus with the near future matters available, the model can amplify it for further futures. With the

improvement of the model, the time period can be given with the help of a well design formula and enough amounts of existing data.

Finally, there is one thing left which cannot be neglected which is important to the running of the model. Since the similarity and contribution will be calculated to determine the appearance and disappearance of either nodes or edges, namely, there ought to be a threshold value to determine if the change needs to be operated.

3.7 Threshold value

As a kind of value that needs to be set based on the existing dataset, it is reasonable to find an optimal value through iterated testing. The value will be randomly select at the beginning, if a certain threshold value performs well with the training dataset, it might have the highest probability to behave well with the latest data of the same field. Besides, it is also advisable to compute the threshold value from the past dataset. Comparatively, such a practice is more demanding, in addition, since the threshold value under different field could be various, therefore, it is hard to find a unified formula to calculate this value. Generally, threshold in the model can be divided into two kinds, status thresholds and rationality thresholds, the former is the sort of thresholds that determines the appearance and disappearance of links and nodes such like probability and contribution, while the latter is set to guarantee the objectivity of the simulation, for instance, nodes with few links might be announced to be eliminated according to the calculation result, since it is not reasonable for a node in spreading to be eliminated, thus, a rationality threshold is essential to control the results under such kind of circumstance for the objectivity of the prediction result.

With all the components introduced above, the dynamic bipartite network model is able to provide a prediction based on the simulation of network changing, which can help the user to have a general idea about the future happenings. Thus, users can take some measures for the potential results.

Chapter 4
Implementation of the network model

In this section, based on the concepts from the chapter introduced above, practical implementations will be demonstrated here. Specifically, information of two aspects will be given, example datasets and concept fulfillments, the former part includes the data source, preprocessing, and related value settings, while the latter part includes the operation functions, algorithms and parameter configurations for the construction of the network model.

From the angle of any new model, there are two key aspects that will arouse user's interest, what kind of dataset or field is suitable for the model, besides, can it provide a good performance. The former part will be introduced in this chapter, and the latter one will be demonstrated in the next.

Since the new model is aiming at the expanding of application, thus, it is impossible to raise examples from all the potential fields here. For the evaluation of the dynamic bipartite network model, two samples will be given based on different datasets and feature matrix. Their background information, dataset processing and feature selection will be introduced in detail separately. Both of the dataset from two examples have some features in common since they are running under the same network model. Besides, two examples are specially selected to show some of the differences between various datasets. Generally, the dataset should be composed of the following components: A matrix that demonstrates the current link status between two groups and a table containing the evaluation of the decisive factor that will influence the changing of the matrix. The model will process the prediction according to the basic information defined ahead. Apart from this, there exist some typical characteristics that could vary from field to field.

4.1 System requirements and hardware developing

- Background introduction

The dataset originates from the relationship between system requirements of different sort of software and the development of computer hardware components. With the estimation of the hardware development, user can take corresponding actions for a better self development.

- Data features

Generally, system requirements of the software are mainly related to CPU, RAM, hard disk space, graphic card, resolution and so on. Different levels of hardware components can be defined as the nodes from one of the groups, for instance, CPU

with the frequency of 2.5GHz and 3.0GHz can be considered as two nodes from the hardware group. But since there are some similarities between them, thus some improvement measures are necessary to decline its negative influence. Specifically, under such a circumstance, if a higher level node is connected with a node from another group, then all of the lower-level nodes should have a link to the certain node. For example, there are four CPU with different level from level 1(slowest) to level 4(fastest), it is obvious that the performance of CPU-3 is closer to CPU-4 compared with CPU-1 under the definition of similarity, therefore, if links of lower level nodes are not constructed, nodes of different levels seems to be totally unrelated, which is inconsistent with the fact. Besides, in this dataset, decisive factor only come from one of the groups, which is relatively easier to compute the potential timing consumption.

The selected data for this theme displays the system requirements of different sorts of software in 2006, 2008 and 2010 separately. The reliability can be guaranteed because all the figures are obtained from the official websites of the software manufactures such like Microsoft, Adobe or EA sports. Intuitively, with the updating of the same kind of software, the system requirements are raising in some of the hardware components. Besides, requirements of different sort of software varies, it depends on its function and display method.

System requirements Software classification	CPU	RAM	Hard disk	Graphic card
Programming platform (Visual Studio)	1GHz	256MB	3.8GB	16-Bit
	1GHz	256MB	4GB	/
	1.6GHz	1GB	3GB	/
Document processing (Microsoft Office)	Pentium 3	128MB	400MB	/
	500MHz	256MB	2GB	/
	500MHz	512MB	3GB	DX9.0c 64MB
2D processing (Adobe Photoshop)	Pentium 3 or 4	384MB	750MB	16-Bit
	1.8GHz	1GB	1GB	OpenGL 2.0
	Pentium 4	1GB	1GB	256MB
3D processing (Maya)	Pentium 3	512MB	450MB	OpenGL
	Pentium 4	2GB	2GB	OpenGL 2.0
	Pentium 4	2GB	10GB	OpenGL
Database (SQL server)	Pentium 3 1GHz	1GB	2GB	VGA
	2GHz	2GB	2GB	/
	2GHz	2GB	2.2GB	/
Animation processing (Adobe Flash)	Pentium 3 800MHz	1GB	710MB	32-Bit
	2GHz	1GB	10GB	OpenGL 2.0
	Pentium 4	512MB	3.5GB	/
Video Processing (Corel VideoStudio)	Pentium 4	512MB	4GB	128MB
	Pentium 4	1GB	1GB	/
	Core 2*1.83GHz	2GB	3GB	256MB

Sports Game (PES)	Pentium 4 1.4GHz	512MB	4.7GB	G6800 128MB
	Pentium 4 3.0GHz	1GB	6.5GB	G6600 256MB
	Core 2*2.0GHz	2GB	8GB	G7900 256MB
Racing Game (Need for speed)	Pentium 4 2.4GHz	1GB	4.7GB	G6800 128MB
	Pentium 4 3.0GHz	1GB	6GB	GT240 512MB
	Core 2 2*1.8GHz	2GB	6.5GB	Dx9.0C 256MB
FPS Game (Tomb Raider Legend or Call of Duty)	Pentium 4 2.0GHz	512MB	9.9GB	G5900 256MB
	Pentium 4 3.0GHz	1GB	8GB	G6600 256MB
	Pentium 2*2.93GHz	3GB	12GB	GT240 1GB

Table 2: Hardware requirements of different sort of software in 2006, 2008 and 2010.

Besides, some other data is also important for both testing and evaluation, this is a table that demonstrates the different levels of computer classified by their price in various years. Specifically, A in the brackets stands for the computer with the lowest performance while D stands for the highest performance. In this sample, DELL is chosen as the target branch for all the computers. The configuration of the latest computers can be found on the official website of DELL, while the other figures can be obtained from the database of some authentic famous websites. Generally, the level of the same configuration is declining with time goes.

Hardware components Series(year)	CPU	RAM	Hard disk	Graphic card
D510(2006A)	Celeron 1.5GHz	256MB	40GB	Intel GMA 900
640M(2006B)	Core2 2*1.66GHz	512MB	120GB	Intel GMA 950
Inspiron 1501(2006C)	AMD 2*1.8GHz	1GB	120GB	ATI X1150 256MB
XPS M1710(2006D)	Core2 2*2.33GHz	2GB	120GB	7950GTX 512MB
OptiPlex 360 (2008A)	Pentium 2*2.5GHz	1GB	250GB	Intel GMA 3100
Inspiron 518 (2008B)	Core2 2*2.66GHz	2GB	500GB	ATI 3450 256MB
Studio (2008C)	Core2 4*2.33GHz	4GB	500GB	9800GT 512MB
Studio XPS(2008D)	Core I7 4*2.66GHz	6GB	1TB	GT220 1GB
VOSTRO 220(2009A)	Pentium 2*2.6GHz	2GB	320GB	Intel GMA X4500
Inspiron 545(2009B)	Core2 2*2.93GHz	2GB	500GB	GT220 1GB
VOSTRO 430(2009C)	Core I5 4*2.66GHz	4GB	320GB	ATI 4350 512MB
Studio XPS 435t(2009D)	Core I7 4*2.66GHz	6GB	1TB	ATI 4670 512MB
Inspiron 560(2010B)	Core2 4*2.5GHz	4GB	640GB	GT310 512MB
Studio XPS 7100(2010C)	AMD 4*2.8GHz	4GB	500GB	ATI 5450 1GB
Inspiron 660-1(2012A)	Pentium 2*2.7GHz	2GB	500GB	Intel GMA
Inspiron 660-2(2012B)	Core I3 2*3.3GHz	2GB	500GB	GT620 1GB
Inspiron 660-3(2012C)	Core I5 4*3.0GHz	4GB	1TB	GT640 1GB
XPS 8500(2012D)	Core I7 4*3.4GHz	4GB	2TB	AMD 7870 2GB

Table 3. Hardware configuration of computers in different years, especially, from 2010, DDR3 is put into use instead of DDR2 in RAM.

Based on the two tables above, a matrix of links between software and hardware requirements of different level can be processed for the simulation. There is one thing need to be paid attention to is that the lowest level of hardware is not placed in the matrix since according to the feature of this dataset, there must be a link between the software and the lowest level hardware. In other words, if the hardware in the matrix is going to be eliminated, it implies that the certain kind of hardware would become the very basic level at that period.

Software\Hardware	CM	CU	CH	RM	RU	RH	DM	DU	DH	GM	GU	GH
Programming platform	1	0	0	1	1	0	1	1	0	0	0	0
Document processing	0	0	0	1	0	0	0	0	0	0	0	0
2D processing	1	0	0	1	1	0	0	0	0	1	0	0
3D processing	1	1	0	1	1	1	1	1	0	1	1	0
Database	1	0	0	1	1	1	1	0	0	0	0	0
Animation processing	1	1	0	1	1	0	1	1	1	1	1	0
Video Processing	1	1	0	1	1	0	1	0	0	1	0	0
Sports Game	1	1	0	1	1	0	1	1	0	1	0	0
Racing Game	1	1	1	1	1	0	1	1	0	1	1	1
FPS Game	1	1	0	1	1	0	1	1	1	1	1	0

C: CPU R: RAM D: DISK G: Graphic card
M: Medium U: Upper middle H: High

Table 4: Matrix of relationship between hardware and software, which is used in the actual simulation of the program, the analysis result is obtained based on this matrix.

4.2 Techniques and application fields

- Background introduction:

This is a relationship network between the computer science technology and their related application field. With the current data available, the dynamic bipartite network model will predict the spreading of the techniques and their extinction time period. From the angle of the users, they are able to set a better business strategy according to the results.

- Data feature

The matrix which demonstrates the relationship between the techniques and its related application field is manually created because of the limitation of data available. But we can still evaluate the result given by the programme via some objective materials. Apart from this, differ from the former sample, there is no relationship between the nodes from the same example, therefore, there is no need to process the whole dataset like the practice mentioned above.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
a	1	0	0	0	0	0	0	0	0	0	0	1	0	1
b	0	0	0	0	0	0	0	1	0	0	0	0	0	0
c	0	0	0	0	0	0	0	0	1	0	0	0	1	0
d	0	1	1	0	0	0	0	0	1	0	1	0	1	1
e	1	0	0	0	1	0	1	0	0	1	0	1	0	1
f	1	1	1	0	0	0	1	0	0	1	1	0	1	1
g	0	1	1	1	1	1	0	1	1	1	1	0	1	1
h	1	0	0	1	0	1	0	1	1	1	1	0	0	1
i	1	0	0	0	1	0	1	0	0	1	0	1	1	1
j	1	1	0	0	0	0	0	0	0	1	0	1	0	0
k	1	1	1	0	1	0	1	0	1	1	0	1	0	1
l	0	0	0	0	0	0	1	0	0	1	0	1	0	0
m	1	0	0	0	0	1	0	0	0	1	0	0	0	0
n	0	1	1	0	0	0	1	0	1	1	0	1	1	1
o	1	0	0	0	1	1	1	0	1	1	0	0	0	0
p	1	1	0	1	1	1	0	1	1	1	1	0	1	1
q	0	1	0	1	1	1	1	0	1	1	1	0	1	1
r	1	1	0	1	0	1	0	1	1	1	1	0	1	1

Table 5. The linking relationship table between techniques and application fields, the relative names of the table will be listed in the appendix.

So far, the preparation of the dataset including the difference between the examples have been introduced in detail, in the following pages, all the necessary data like the feature list for model execution will be configured based on the introduction of the construction principles.

4.3 Construction principle

Briefly speaking, dynamic bipartite network model is composed of the appearance of the links between two nodes from two groups and extinction of a specific node from one of the groups. The former is determined by the probability between two nodes, while the latter is decided by the contribution of one node. Moreover, the result is not simply calculated from the original data matrix, it will give the output after the upgrade of the matrix for a couple of times, which stands for the imitation of the

dynamic development of reality. Generally, the processing flow of the dynamic bipartite network can be demonstrated through the figure below.

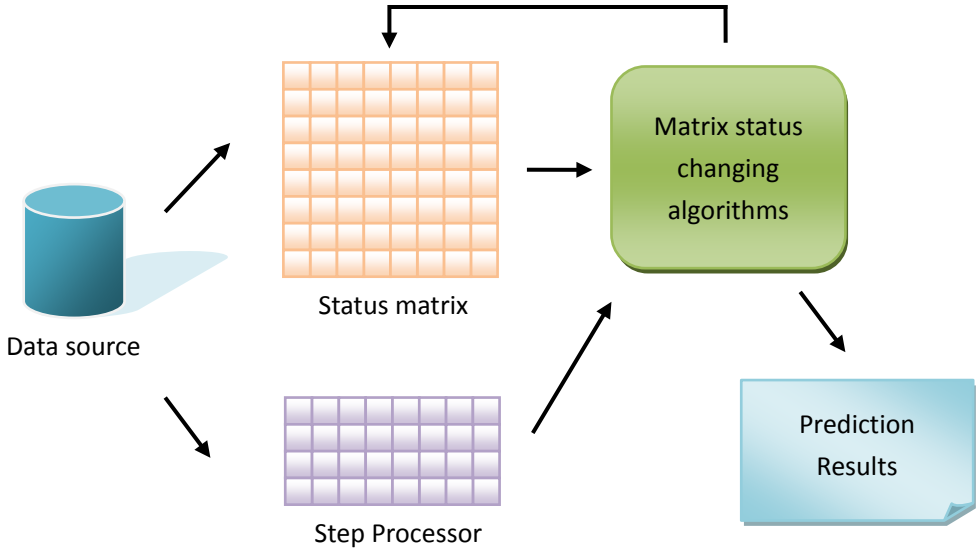


Figure 11. The operation flow figure of the basic dynamic bipartite network, which contains the major procedure and the main components in use as well.

From a macro view, the network model is consisted of two main components: data and algorithms. Data from the selected field will be preprocessed to an original status matrix and a step processor as well. Based on different way of implementing the step processor, the result will be different based on the same input. With the assistant of the step processor, the algorithms get the current state from the status matrix and change the status of the matrix according to the judgment rules. Such a process can be iterated for certain rounds based on the needs of the users. To a microcosmic view, there are some components which are important for the construction of the model will be briefly introduced as follow:

- A matrix processor, which can store the information of the linking status between groups, most important of all, it should have the function of matrix modification such like value changing, column deleting, variation recording and other related functions.
- Mathematical processor, since the main purpose of the basic model is to calculate the probability and contribution of the matrix, therefore, related function is necessary, such like the random column creating and computing of mean and standard deviation, which are both of great significance for the calculation of the contribution. On the other hand, this component should work well together with the matrix processor.

From the introduction of the construction principle especially the flow graph, it is not difficult to find that the most important component to distinguish the difference between bipartite network and dynamic bipartite network is the construction of step

processor. Therefore, both of the time evaluation method will be introduced here based on the real data.

4.3.1 Feature list setting

- System requirements and hardware developing:

In this sample, node features are only chosen from one of the groups, in other words, it is assumed that only some features from the hardware components will influence the popularisation. Specifically, price is an important parameter that cannot be ignored, except for a small part of users who do not care about that, the majority of the users are always prefer to purchase a high cost performance product. Intuitively, a high performance product is hard to be popular unless its price is declined. Besides, another feature of the hardware components also influences the popularisation to a high extent. Specifically, we define this feature "efficiency", to make it clearer, taking CPU and hard disk as an example, the function for hard disk is just to store all the program codes. There is no difference to install a piece of software in a hard disk with a bigger or smaller space. As to the CPU, when running a demanding software, a faster CPU will process the operations in a shorter time than a slower CPU, which means CPU is more related with "efficiency" than hard disk. Comparatively, component with a higher "efficiency" probably has a faster updating speed than a lower one.

Feature\Hardware	CM	CU	CH	RM	RU	RH	DM	DU	DH	GM	GU	GH
Cost	4	6	10	1	2	3	1	2	3	4	6	10
Efficiency	3	3	3	2	2	2	1	1	1	3	3	3

Table 6. The table of the features of the matrix with the fixed specific values.

Based on the feature list, the way of timing consumption evaluation can be calculated through the method based on the idea of weight. Directly, the evaluation formula can be written as follow:

$$Time\ Consumption = \frac{\sum_{1}^{n} Value_n \times Weight_n}{Timing\ Base}$$

To be more specific, n in the formula stands for the total number of the features in the feature list. Weight can be considered as a kind of importance tendency, the sum of weight of all the features should be a constant value such like 10, 20 or 100. Timing base is a parameter to control the pace of matrix changing. In the sample testing, in order to have a reasonable comparison with the method based on Perez's theory, the sum of weight of features is 10, cost weight 6 and efficient weight 4, timing base is set as 7, in order to have same maximum time consumption with the method about Perez's curve.

- Techniques and application fields:

In this relationship network, features will come from both of the groups. From the technique side, setup cost is been extracted as a typical feature, it reveals the level of comprehensive difficulty of applying a specific technology to a certain field, including expenses, potential manual work amount and so on. On the other side, “interaction” extent is chosen from the application field. It stands for the level of interaction between individualities from the certain field. For instance, if a new technique of engine is going to be applied on the taxis, the cars either with old engine or new engine can be on the run fluently. However, if a new flash technique is applied on a website, all the users have to download the plug-ins to see the special visual effect. Thus, it seems to have a higher “interaction” level than the engine technique. In general, it seems that techniques will spread slower in the field which has a high “interaction” level. Finally, the scope of the application field is applied into the feature list. This is a criterion that evaluates the amount of users, intuitively, the bigger the scope is, the slower the spreading speed will be.

Technique	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**	**K**	**L**	**M**	**N**
Difficulty	3	2	3	2	3	2	2	2	3	2	1	1	1	1

Field	**a**	**b**	**c**	**d**	**e**	**f**	**g**	**h**	**i**	**j**	**k**	**l**	**m**	**n**	**o**	**p**	**q**	**r**
Interaction	0	0	0	0	1	1	0	1	1	1	1	1	1	1	1	0	1	0
Scope	4	2	2	3	4	3	2	4	4	4	3	3	2	3	2	1	1	1

Table 7. The feature list of the relationship network between techniques and fields, the relative names can be seen in the appendix.

4.3.2 Perez’s curve simulation

Apart from the feature list, the simulation based on Perez’s curve is also a sort of solution. Specifically, a step table is set to search for the specific steps in the different period of the technique development. Since the curve has the shape similar to the tangent curve, hence, time consuming for a link during the beginning and ending is comparatively slower than in the middle.

Timing consumption based on Perez’s theory has several methods for implementation, different methods can be applied under different scale of the dataset or the precision demand according to the fact. For the efficient method, a table that evaluates the time consumption in different period can be set beforehand. Such a practice is used in the testing of the model.

Period index	**0**	**1**	**2**	**3**	**4**	**5**	**6**
Time consumption	7	4	2	1	2	4	7

Table 8. The table of step settings according to the Perez’s curve, intuitively, figures in the table are symmetrical.

The construction principle will be explained at first, period index can be considered as the current period of the development of a certain object, which is influenced by the amount of current links with others. According to the concept of Perez's theory, the developing speed is fast in the middle and slow in both of the ends. Thus, the index in the middle has the shortest time consumption to construct the link in the data matrix. From the angle of mathematics, if the number of steps presents for the time consumption, its reciprocal stands for the speed of the construction. Based on the data from the table, if we combine some straight lines with the slope from 1/7, 1/4, 1/2, 1 and the reverse sequence back to 1/7, a rough polyline with a shape similar to Perez's curve is built for the further calculation. Intuitively, if more period indexes are set in the table, the time consumption evaluation will be more precise. From the other hand, such a design can be applied to the data with all sorts of scale. The formula of time evaluation can be written as follow:

$$Time\ Consumption = Table(\frac{Current\ Connections}{Maximum\ Connections} \times Maximum\ Index\ )$$

Raising an example with random numbers, if the amount of the possible links between a certain node and all the nodes from the other group is 40, while there exist 24 edges at the moment, based on the table given above, the index for the current situation will be 4 with rounding, thus, the time for edge construction will be 2.

4.3.3 Rationality thresholds configuration

In order to have a more objective prediction result, there are some other parameters being set in the model. Disappearance delay is one of the parameter which is a kind of assistant for the model, from the view of the reality, there might be a gap between the coming of the new node and the disappearance, in other words, both of the old technology and the new technology will coexist for a short period of time. Such a phenomenon can also be seen from the figure of Perez's theory. In the real data testing, time delay is set for 2. When a certain node is judged to be eliminated, the actual disappearance of the node will be two steps afterwards.

Besides, there is a special parameter which is set to avoid the potential problem during the calculation of the contribution. According to the definition of contribution, under the situation that all the random permutation of a certain column will be displayed, the nestedness of the matrix needs to be calculated for the further computation.

According to the definition of "discrepancy" method, both of the groups will take only one step to complete the transformation. Besides, the sum of nestedness from the random column permutation will also be the same, therefore, both of the matrixes will have a same value of contribution. But obviously, the situations from these two groups are absolutely different. Based on the research of contribution, it implies that the node which has more links with others have a higher probability to be eliminated, thus, in order to predict the right node that will face the extinction, a simple but useful

mechanism will be applied into it. Apart from checking the contribution of the certain node, the proportion of the edges will be inspected before making a prediction. In other words, if the proportion of current edges which have been built of the node does not reach a threshold, even if the contribution is high enough, it still needs more links for development. In the real testing, the proportion is setting for 60%.

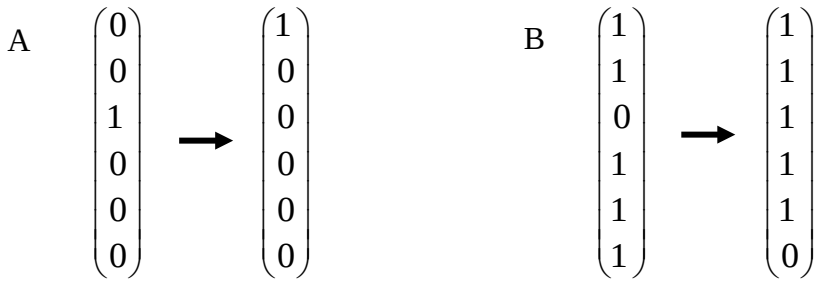


Figure 12. The process of matrix transformation based on the “discrepancy” method of nestedness calculation.

So far, the related concepts of the model, their implementation algorithms, data characteristics, parameter configurations and sample data information have all been specifically introduced, in other words, a model of dynamic bipartite network can be completely constructed based on the details mentioned above. Therefore, the result evaluation will be demonstrated in the following chapter.

Chapter 5
Analysis of prediction results

Since the major purpose of the new model is prediction, thus, in order to draw an objective evaluation of the accuracy of this model, the source of the dataset should be convincing. The whole data for model evaluation is made up of two sections. From one hand, the figures of the tables for development simulation ought to be correct, otherwise, any prediction based on a suspected dataset will be regarded useless. From the other, the actual developing data is of great significance, via comparing the predictions and the actual result, a conclusion of the evaluation of the new model can be obtained, which would help for the improvement of the model.

In the basic version of the dynamic network model, the concept of steps is used instead of time because of the data limitation. Therefore, the accuracy of developing sequence is the main method for model evaluation. Besides, simulation of the life span of a node is another reasonable method, still the time period of one step is necessary for this method.

Evaluation of the basic dynamic bipartite network model can be divided into several sections which cover the following aspects. In each aspect, different performance of the model with different threshold will be demonstrated, in order to bring a more comprehensive evaluation of the model.

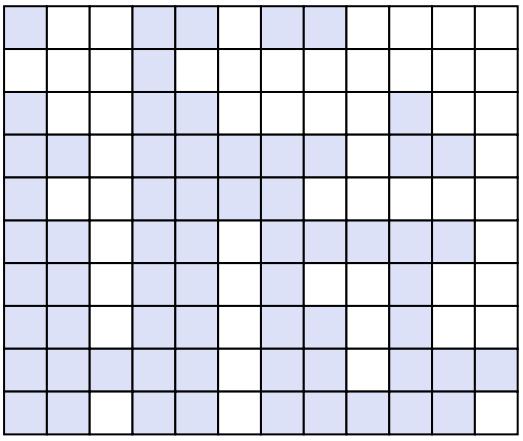
- Sequence accuracy: Since there are no random values in this model, thus the sequence obtained from the model is constant under the same situation. Via comparing the result with the actual development in the selected area, a sequence difference can be obtained. From the angle of the model, sequence is a key measurement criterion because it would help users to have a direct feeling of the future happenings.
- Life span accuracy: This is another important criterion for results evaluation, based on the features of the model construction principles, the main method for life span analysis is to comparing the multiple relationships between the time spot of extinction of different nodes.
- Perez’s curve analysis: Such an analysis aims at the changing pace for one node, it focuses on the phenomenon that if the selected node is expanding its number of links like the tendency according to Perez’s curve.
- Special theories: In different areas, there are some special theories which can partly reveal the accuracy of the model. This can also be a kind of assistant method.
- Timing consumption: Efficiency is a criterion that we cannot neglect, in this

section, different algorithms will be used to run the model only for the difference of time, for a good solution, high accuracy and fast speed are both demanding.

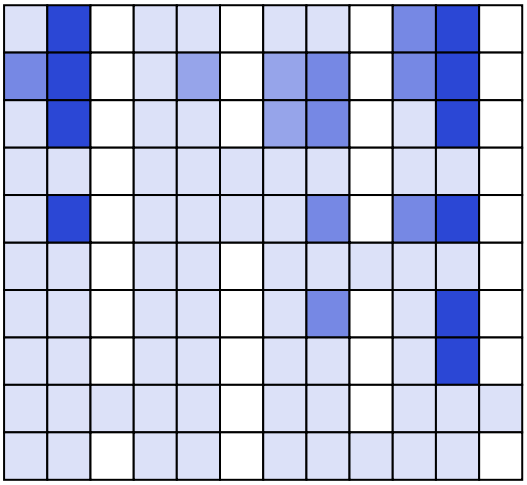
Apart from the analysis methods, all the assumptions and value configuration is ought to be given before the demonstration of the results calculated by the network model. In order to have an objective result for analysis, on the general level, different threshold values are in use to see the difference of the output, in addition, both of the feature list and Perez's curve are applied for a comparison. Since the output of the program is not direct enough, thus, the matrix graph will be used for the demonstration of the changing of the matrix. The darker the dot in the matrix, the later the link between nodes is constructed. Since there are some limitations of the data source, hence, the relationship matrix of techniques and applications will be added into the analysis which is not closely related to the actual result comparison. From the angle of the matrix development, it is assumed that the information from the matrix is a kind of reference for the successors, they will determine their actions based on the current information. Raising the relationship between hardware and software as an example, if a programmer who only use the computer for programming tries to make a decision of the hardware component for use, he will analyse the current information to find the similarity with others, and determine if a certain kind of hardware component is necessary to update. Besides, no new coming nodes is used in the testing, in other words, although the information of node extinction will be displayed, in order to have an entire view of the whole matrix development, the nodes will not actually be deleted for sample testing.

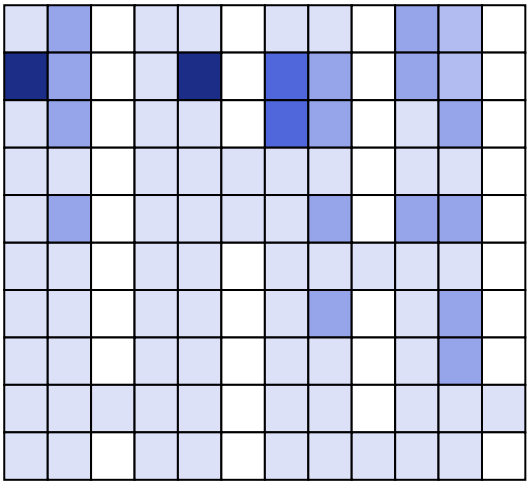
In the real testing, the following threshold pairs of probability and contribution are applied into the data analysis: (0.3, 0.5), (0.3, 0.8), (0.3, 1.2), (0.3, 2.0), (0.5, 0.5), (0.5, 0.8), (0.5, 1.2), (0.5, 2.0), (0.7, 0.5), (0.7, 0.8), (0.7, 1.2), (0.7, 2.0), (0.8, 0.5), (0.8, 0.8), (0.8, 1.2), (0.8, 2.0). Since no new coming nodes are applied in the testing, hence, the appearance and disappearance of the edges and nodes are only related to the similarity and contribution. Namely, with 16 pairs of testing value, only four kinds of development will be presented, and four sorts of nodes disappearance are shown.

Original matrix

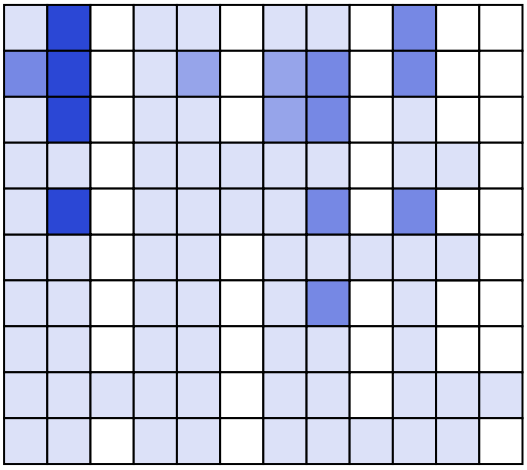


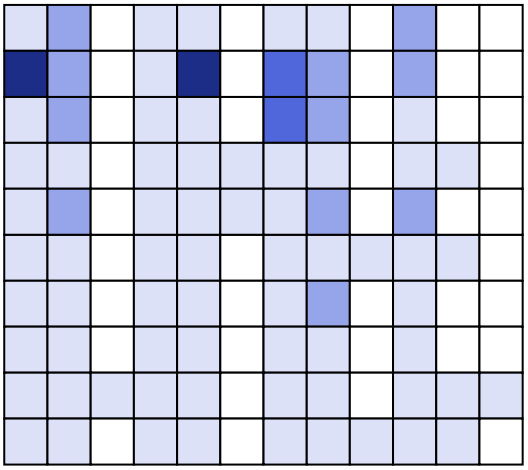
Group A



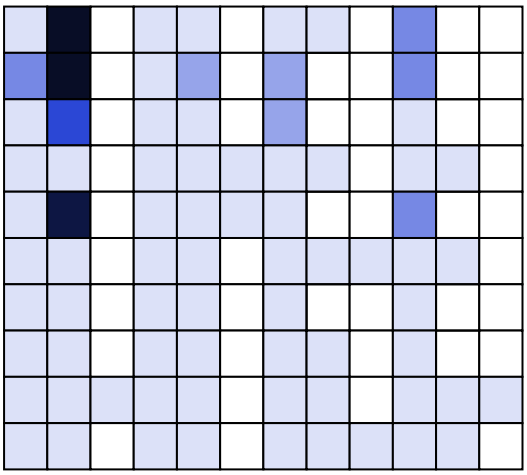


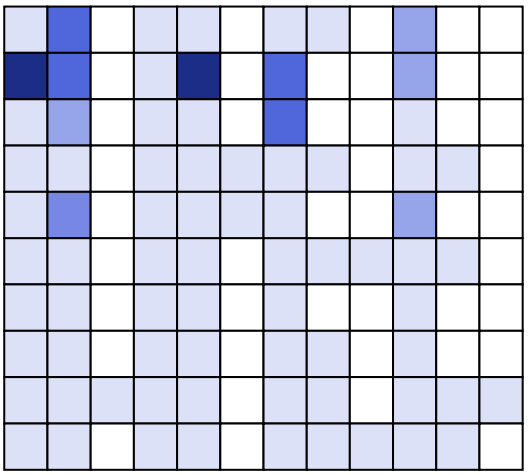
Group B





Group C





Group D

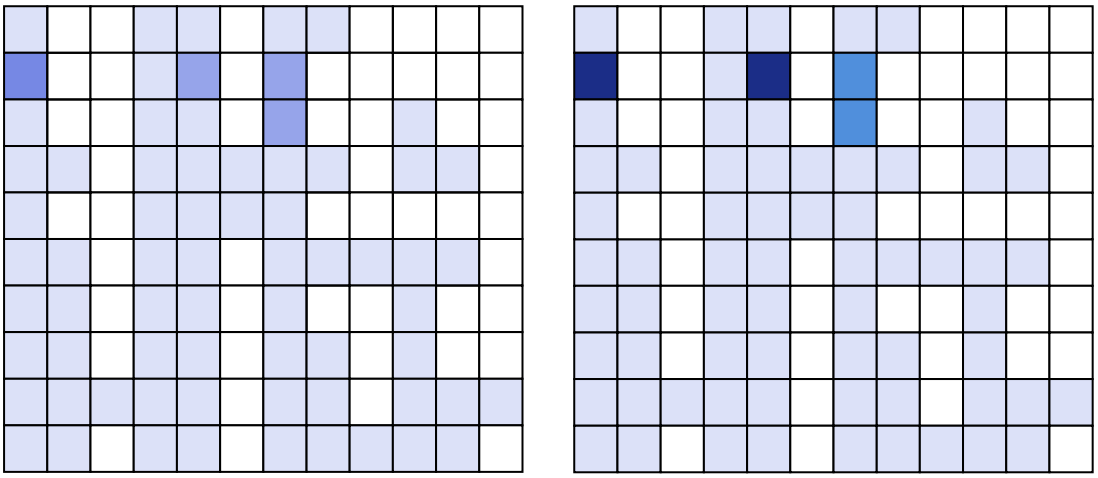


Figure 13. The comparison of different probability threshold and between two timing computing methods. Group A, B, C, D represent the probability of 0.3, 0.5, 0.7 and 0.8 separately.

Based on the information revealed from the matrix graph, some direct conclusions can be made at first. Obviously, the number of link construction declines with the increase of probability. 21 new links are built when probability is 0.3, on the contrast, only 4 new edges are built when it is raised to 0.8. Since no nodes are deleted in the testing, thus, the matrix developing based on either feature list or Perez's curve provides the same result in the end, the only difference is the sequence of appearance. Besides, one feature of the result also reveals a demanding of the original data set. Specifically, if a node has only small number of links with the nodes from the other group, nodes which originally have 1 or 2 links do not have the chance to establish new links even if the probability is set on 0.3. Thus, improvements should be designed for such a circumstance.

Table A

	a	b	c	d	a	b	c	d	a	b	c	d	a	b	c	d
CM	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6
CU	8	8	8		8	8	8		8	8	8		8	8	13	
CH																
RM	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
RU	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
RH																
DM	3	3	3	3	5	5	5	5	5	5	5	5	5	5	5	5
DU	6	6			6	6			6	6			6	6		
D																
GM	3	3	3	3	3	3	3	3	3	3	3	3	6	6	6	
GU	8				8				8				8			
GH																

Table B

	a	b	c	d	a	b	c	d	a	b	c	d	a	b	c	d
CM	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10
CU	5	5	5		5	5	5		5	5	5		5	5	7	
CH																
RM	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
RU	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10
RH																
DM	3	3	3	3	7	7	7	7	7	7	7	7	7	7	7	7
DU	5	5			5	5			5	5			5	5		
D																
GM	3	3	3	3	3	3	3	3	3	3	3	3	5	5	5	
GU	5				5				5				5			
GH																

C: CPU R: RAM D: DISK G: Graphic card
M: Medium U: Upper middle H: High

Table 9. The result of disappearance of the nodes predicted by the network model, figures in the table represents the time spot that a certain node is announced to be out of date. a, b, c and d represent the probability threshold value which is set for 0.3, 0.5, 0.7, 0.8 separately. Four sets of figures from the same table show the change of disappearance under different value of contribution, specifically, 0.5, 0.8, 1.2, 2.0 from left to right, which is divided by a coarse line. In addition, table A is obtained from the model based on feature list, while table B is based on Perez's curve.

Some general information can be seen from the tables. First of all, under the same value of contribution, the phenomenon of node disappearance declines with the increase of the probability, specifically, the prediction of the disappearance of the higher level hardware is cancelled by the model. Besides, the increase of contribution leads to the delay of the node disappearance. Some of the node disappearance is postponed with different extent, some of the time consumption is even doubled. With the intuitive comprehension of the result, the following analysis will be described in detail via the methods which has been mentioned above.

5.1 Sequence and life span analysis

On the view of hardware development, old techniques will be substituted by the latest ones can be regarded as a kind of law. From the result of the testing data, predictions based on the feature list perform well since every higher level component is predicted to be "extinct" after the disappearance of the lower level component. On the contrast, the prediction based on Perez's theory provides a not good enough result. Specifically, sequences in CPU and hard disk are reversed either because of the timing setting theory or the changing of contribution. On the other hand, sequence between

components is also an important criterion, which will be analysed based on the actual data demonstrating as follow.

Hardware	Eliminated Time	Hardware	Eliminated Time
CM	2009	DM	M-2012
CU	M-2010	DU	M-2012
CH	U-2010	DH	U- 2012
RM	2010	GM	2009
RU	M-2012	GU	2012
RH	U-2012	GH	U-2010, M-2012

Table 10. The elimination of the hardware components of the year 2008. Some of the components are still in use at the present thus, the latest level is labeled.

From the information of the table based on the real data, it is evident that CPU and graphic card of medium level in 2008 is the fastest to be eliminated, higher level of these two components are also going to the lower level in the following years. Comparatively, RAM is eliminated slower, although DDR2 is out of date since 2010, from the angle of performance, they can still occupy the market for a certain period. Intuitively, hard disk is the most long-life component, even the capacity of 2TB appears in 2012, 500GB and 1TB hard disk is still in use in the medium level and upper middle level separately. From such an analysis, the sequence of updating frequency between hardware components can be listed as CPU, graphic card, RAM and hard disk from the fastest to the slowest. As to the prediction results, it seems that the network model is sensitive to the nodes which is widely linked, such a phenomenon is normal based on the principle of the model construction. Aiming at the specific data, RAM and hard disk are comparatively slower to be eliminated than the graphic card, but the sequence of CPU is not in the right position, from the angle of machine learning, it is not reasonable to judge the quality of a predictor based on the performance of one dataset, but at the same time, the example reveals that it is of great importance to find some proper criterions in the feature list and set the related values reasonably.

As for the life span, it can be obtained from the testing data that when the contribution threshold is comparatively higher, the time spot of disappearance seems to be closer. Since the original data is collected in 2008, thus it is possible to find the timing relationship between nodes. From the actual results, the prediction with a middle lower contribution and feature list is most accurate to the fact. Specifically, there is a gap of one year between the position changing of the medium and upper middle CPU, hence, the actual timing difference is doubled, and 6:8 is demonstrated in the result. Similarly, 1:2 in RAM, and 3:5 is obtained, 1:1 in hard disk, 5:6 is obtained, and 1:4 in graphic card, 3:8 is obtained. Such a result is based on the parameter of 2 step delay of the node disappearance, if the parameter is step to 1 step, the actual result will become 5:7, 2:4, 4:5 and 2:7, which is much closer to the real data. On the whole, the model provides a reasonable result of the timing multiple relationship between nodes.

Similar to the conclusion of the sequence analysis, in order to improve the quality of the prediction results, it is significant to construct a current dataset of a high quality.

5.2 Perez's curve analysis

In the accuracy analysis, it is obvious that the performance based on Perez's theory is comparatively inferior to the version with a feature list. In this part of analysis, a total different angle will be displayed, according to Perez's theory, a certain technology will develop fast during the middle of the whole period, while during the beginning and end, the speed is comparatively slower. Focusing on the developing matrix graph demonstrated above, there is an evident feature that links in one node are almost constructed in the same round, which is not objective enough in reality. From the view of the data construction, since there are not enough parameters in the feature list to distinguish the difference between nodes, it is probable to have such a prediction especially the amount of feature list is low from one group.

Based on the performance of various parameter pairs, according to the performance of the execution results, the model is running again with fixed parameter thresholds and different dataset for a second time evaluation, whose prediction results are displayed as follow:

	Feature list	Perez's curve
CM	6	10
CU	/	/
CH	/	/
RM	3	3
RU	5	7
RH	6	5
DM	5	10
DU	6	10
DH	6	10
GM	3	3
GU	8	5
GH	/	/

Table 11. The result of prediction with the probability of 0.3 and the contribution of 0.8 based on the matrix of 2006 of the relationship between software and hardware.

From the angle of sequence analysis, the method of feature list still provide a better performance than Perez's curve function, specifically, all the elimination sequence of hardware components are still will displayed from feature list, while sequence of Perez's curve have a little mistake in RAM.

Hardware	**Eliminated Time**	**Hardware**	**Eliminated Time**
CM	2007	DM	2007
CU	2007	DU	2007

CH	M-2008	DH	2007
RM	2007	GM	2007
RU	2008	GU	M-2008, 2009
RH	M-2008	GH	U-2008, M-2009

Table 12. The elimination of the hardware components of the year 2006.

As for the life span analysis, the accuracy of feature list still stands at a high level which can be displayed via a table as follow:

Hardware\Elimination	Reality	Prediction
CPU	1,1,2	1,/,/
RAM	1,2,2	3,5,6
Hard Disk	1,1,1	5,6,6
Graphic Card	1,3,4	3,8,/

Table 13. Difference between the results from reality and prediction based on the dataset of 2006.

Specifically, the ratio of each hardware components from the prediction seems close (almost 100%) to the reality exception for the missing of a little proportion of the predictions (25% in this sample). Thus, from an intuitive but also objective view, such a conclusion can be made that the dynamic bipartite network can be used for the work of prediction, threshold configurations can be decided by the learning of past data. Besides, values in the feature list are demanding since it will directly influence the result of the final prediction.

For another sort of case, a dataset which is assisted via a list of 3 features and bigger scale, something different will be displayed:

	A	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**	**K**	**L**	**M**	**N**
a	1	6	0	0	8	11	6	0	8	6	8	1	4	1
b	6	5	0	5	6	5	11	1	6	5	3	0	3	3
c	6	5	0	5	6	5	11	0	1	5	3	0	1	3
d	7	1	1	9	12	9	11	0	1	5	1	0	1	1
e	1	7	0	0	1	12	1	0	9	1	10	1	8	1
f	1	1	1	0	13	10	1	0	8	1	1	0	1	1
g	6	1	1	1	1	1	11	1	1	1	1	0	1	1
h	1	7	0	1	14	1	13	1	1	1	1	0	5	1
i	1	7	0	0	1	15	1	0	9	1	10	1	1	1
j	1	1	0	0	14	12	7	0	9	1	9	1	5	5
k	1	1	1	0	1	14	1	0	1	1	10	1	5	1
l	8	6	0	0	8	14	1	0	8	1	10	1	5	5
m	1	6	0	0	7	1	6	0	7	1	4	0	4	4
n	8	1	1	0	13	14	1	0	1	1	8	1	1	1

o	1	6	0	0	1	1	1	0	1	1	7	0	4	4
p	1	1	0	1	1	1	9	1	1	1	1	0	1	1
q	7	1	0	1	1	1	1	0	1	1	1	0	1	1
r	1	1	0	1	11	1	9	1	1	1	1	0	1	1

Table 14: The status matrix based on the relationship network between technology and application, 0.5 and 2.0 are set for probability and contribution separately. 1 in the matrix means that the link is originally built in the matrix, while 0 reveals the unconnected situation between nodes. Besides, other numbers in the matrix demonstrate the link construction sequence of the matrix, the bigger the number is, the later the link appears.

From the result table showed above, the appearance sequence becomes asynchronous, comparing with the dataset of hardware and software. It seems that the increase of feature amount will help to amplify the difference between nodes. Besides, the enhance of scale of the matrix will also improve the diversity since more element will take part in the calculation of the probability. In addition, it is not suggested to list too many features for value computing, on one hand, it will cost the computing efficiency, on the other, it is not reasonable to list too many features that is closely related to the prediction result.

5.3 Related theory analysis

Apart from the self data checking, some existing rules are also useful for result evaluation, such like Moore's law [12]. This is a classical law in the area of hardware development, it is constructed based on the observation of the history of computing hardware. The main context of the law is that within the same purchasing ability, the number of transistors on the integrated circuits will be doubled in eighteen months. Intuitively, such an assumption will not always be correct. In fact, with the improvement of hardware processing, the law might lose its significance in the decade, but it can still work at the moment.

From the angle of result evaluation, although the fastest CPU is developing at a similar speed based on Moore's law, it is impossible to apply the latest hardware technology to every level of the computer. Therefore, it is more reasonable to analyse the level reduction of the same CPU since the latest technique will be used on the top level computer. Focusing on the real data, the CPU of the medium level computer varied from dual core in 2008 to quad-core in 2010, which is used in the upper middle level in 2008, hence, since the result based on 1 step delay can provide a prediction result that the medium level CPU will be eliminated after 4 steps, and upper middle level will be eliminated after another 3 steps, it implies that the elimination frequency of the CPU predicted by the model is close to the fact.

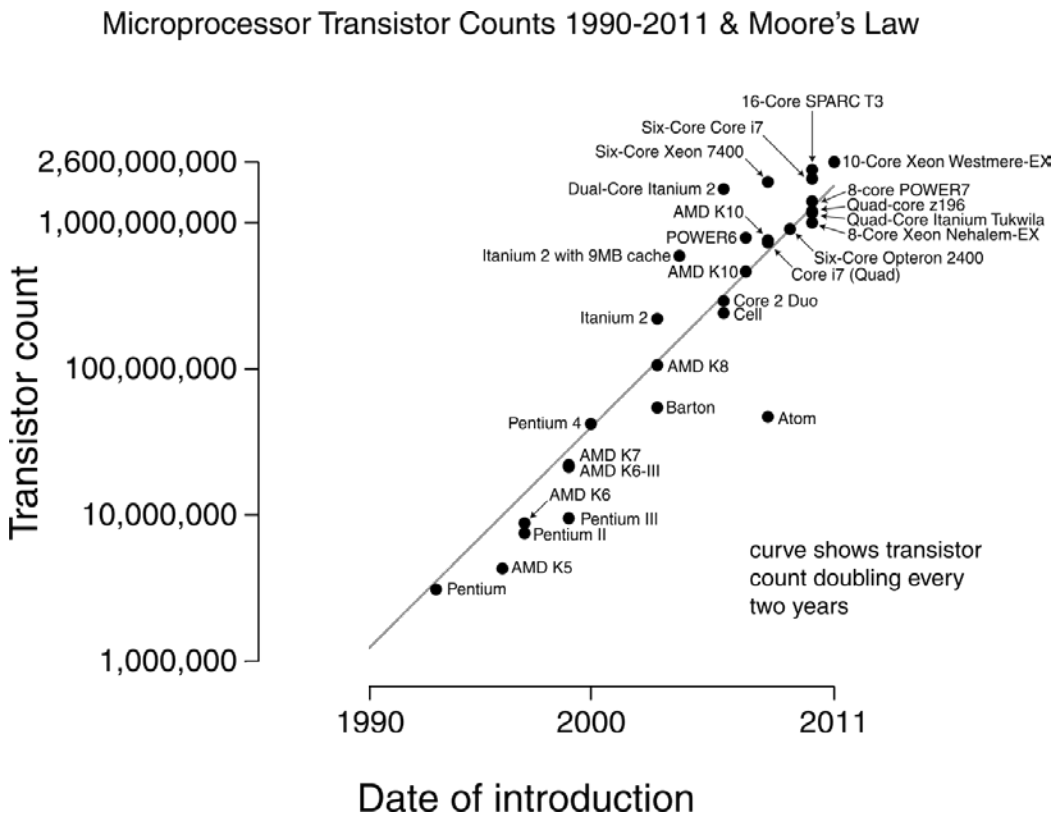


Figure 14. The processor development based on Moore's law. Figure 14 is adapted from [12].

5.4 Time consumption analysis

In this part of analysis, time consumption based on different level of data scale will be demonstrated. The execution time of both of the examples will be listed since they have the different scale of dataset.

Row	Column	Round	Time(s)
10	12	8	0.209
18	14	18	51.974

Table 15. The time consumption of different data scale, specially, the value of round is the final round that the model provides a related prediction, afterwards, the matrix will keep stable.

This is an astonishing result that there is a big difference between a matrix with a maximum of 120 edges and 252 edges. The smaller matrix cost about 0.025 seconds per round, while the bigger one cost about 2.9 seconds. Much more than the difference that shows a double relation between the matrix scales. Such a huge time difference is due to the computing of contribution. Specifically, based on its definition, since it need to list all the combination possibilities, the time consumption will cost a lot with the expanding of the dataset scale. Such a conclusion can be proved by the fact that the multiple relationship between the time consumption is close to their related maximum combination amount. Since the bipartite network originates from

the dataset with the scale of hundred, thus it is significant to have an evaluation of the probable time consumption of dataset with different scale.

N	N!	N! / (M!*M!)
10	3628800	252
18	6.402e+15	48620
20	2.433e+18	184756
50	3.041e+64	126410606437752
100	9.333e+157	1.090e+29
200	7.887e+374	9.055e+58

Table 16. Different N, factorial of N, and number of combination of M from N, while M = N/2, namely, the third column is the biggest number of combination of N.

Obviously, the figure of combination increases with an unbelievable speed with the scale expanding, under such a circumstance, it will cost hundreds of years to compute the table of the scale in hundred. Hence, it is of great significance to enhance the computing efficiency to a new level, otherwise, it is hard to apply the dynamic bipartite network model to the new fields with comparatively big datasets.

Some mechanisms can be used to solve this problem at different extent. Preprocessing is one of the solutions to this situation, since the result of contribution is fixed under the same circumstance, thus, they can be pre-computed via some supercomputers. Besides, other methods without listing all the potential combination of edges can be tried to substitute the current method, however, its correctness and objectivity need to be improved on a mathematical level.

5.5 Conclusion

As a summary of data analysis, several conclusions can be obtained from the results predicted by the network model. First of all, model with feature list has a better performance than that with Perez's theory on the accuracy aspect. Besides, more parameters in the feature list will provide a better performance, since it can distinguish the difference between nodes to a higher extent. In addition, the value configuration of the feature list is of great importance. Finally, some of the parameter setting needs the past data training in order to have a most ideal result among all the possible values.

Chapter 6
Improvements and further work

In this section, improvements and further work will be introduced from three aspects. First of all, some improvements are designed to solve the problem from the results of the data analysis. Besides, some solutions are designed either for the solving of potential problems and risks or for its overall performance enhancement. Finally, examples of cross field application of the dynamic bipartite network model will be introduced for a kind of reference.

6.1 Improvements based on the analysis results

- Self-developing mechanism

In the data analysis, it is revealed that if the amount of edges with one node is not enough, it will keep such a state to the end of the simulation, which does not fit the facts. In order to improve the quality of the model, a self-developing mechanism can be applied into the model. The principle of the mechanism is not difficult, a probe will be set to detect the development of a certain node. For instance, if the probe gets the information that there is not any link construction in a certain period of time, a special algorithm will be activated to build some link without observing the state of the matrix. Such a mechanism has various ways for implementation, first of all, the easiest way to realise it is randomly choose a node from the other group for linking. Besides, it is more reasonable to decide the nodes for linking based on some of the information from the feature list, for instance, links of the hardware can be made based on the purchasing ability of the other group. Apart from this, a dynamic probability threshold is also a kind of solution, namely, the threshold will change based on the current amount of the links, the threshold will be declined if the total number of edges of a certain node is low. At the same time, such a practice might control the node development to follow the form of Perez's curve. Overall, under such a mechanism, every node in the group will be eliminated, the only difference between nodes is the life span.

In the previous samples, the link status between groups is the only thing that we focus on, although the simple version can provide some prediction result, it is obvious that there exist some improvements from different aspects that can refine the whole model, which can either raise its accuracy or expand its application field.

- Algorithm improvements

Although there is not a high demanding for the efficiency of the whole model, especially, the scale of the dataset is not that huge. When the model meet the dataset with a huge scale such like the posts from a social network platform, it is necessary to

enhance the efficiency of the model. Except for the long term prediction, if the model is applied into the online application, it is improper to keep the users waiting because of the limitation of computing algorithm. Therefore, it is significant to find a new system which is compatible with a big scale of dataset.

One of the algorithm will be introduced here as a kind of reference for performance improvement. Assume that both of the groups have n nodes, thus, based on the functions of such a normal matrix, the time complexity for the function dealing with the matrix will be O(n*n). Such a complexity will be very huge if n is big enough, such like the level to million. Under this circumstance, Van Emde Boas tree [13] will be applied to reduce the complexity to O(n*log log n). The “discrepancy” calculation of nestedness is closely related to the maximum and minimum position of 0 and 1 in a specific column. Since vEB tree stores the position information in each of the parent node, thus there is no need to compute the steps of column change in the linear time of O(n).

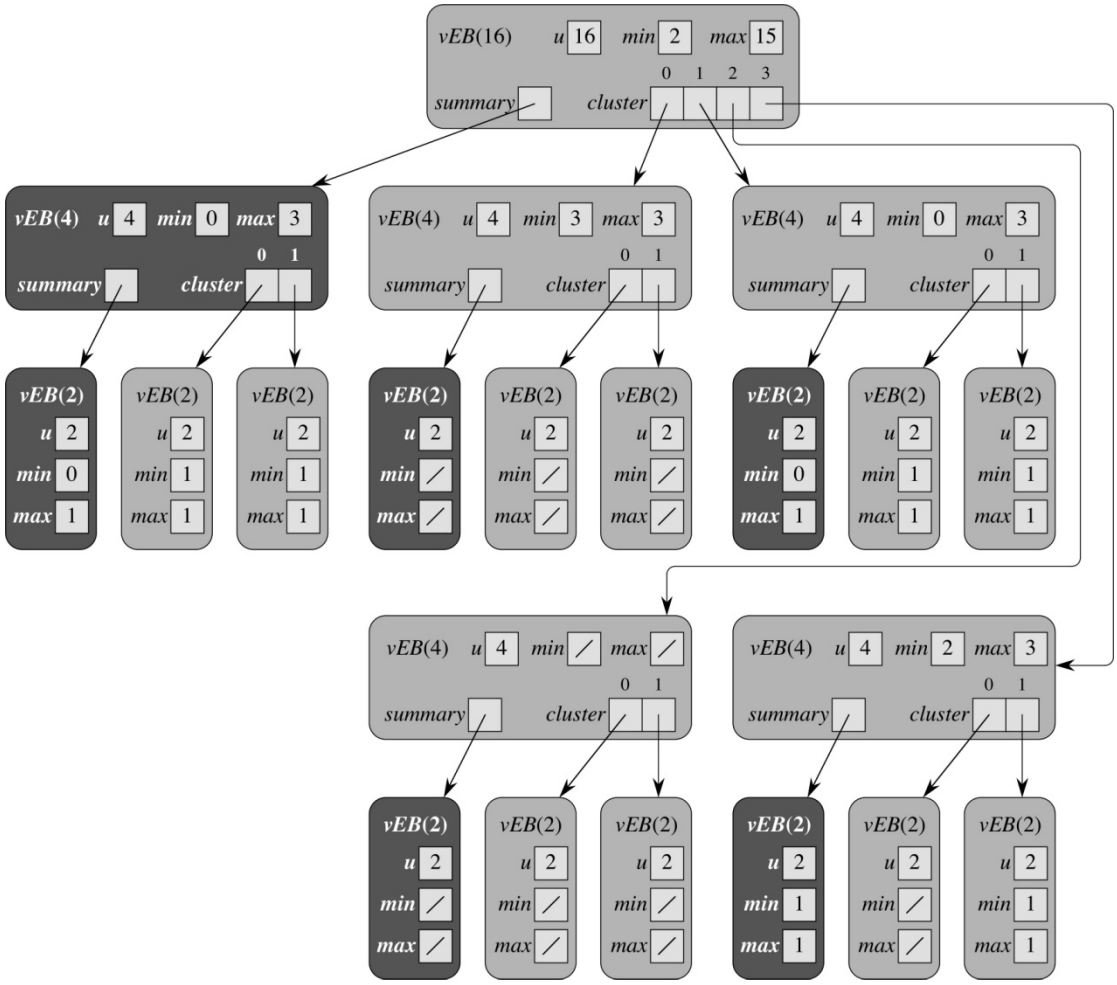


Figure 15. The construction of Van Emde Boas tree. Figure 15 is cited from [13].

Obviously, the original vEB tree only stores the position information of 1, but in the variety which serves for the nestedness computation, information of both 0 and 1

should be stored in the parent nodes since the computation is related to the position swap between them. Specific performance need to be further improved although it seems to be compatible from the theoretic level.

6.2 Future improvements

- Weighted dynamic bipartite network [14]

It will become a totally new version if weighted network is applied into this model. More information can be revealed from it. Since we can add numbers onto the nodes and edges, evidently, concept like scale and allocation will be added into the model. Take technology and its application field as an example, when the dataset is put into the new model, clearly we can see that which technology is more widely spread, and how one specific technology is applied in different areas with different situation. Such a comparison can be demonstrated from the following figure.

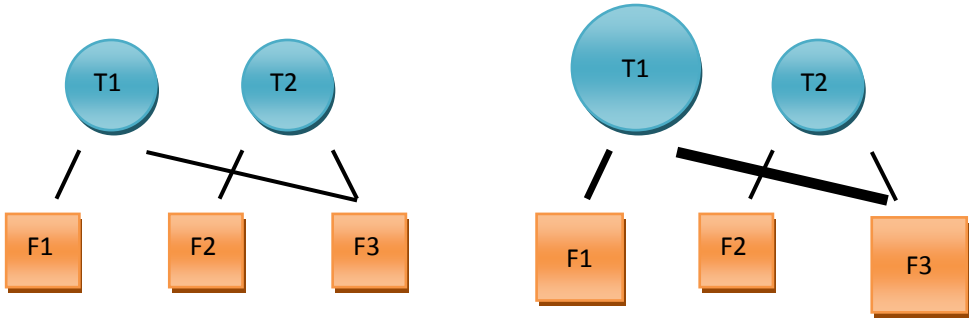


Figure 16. The comparison graph displays a direct new view of the network model. There is not any difference between nodes and edges from the left, but while applied weight into it, information amount is ascending to a new level. Specifically, technology 1(T1) has more variations than technology 2(T2), T1 is more widely applied in field 3(F3) than in field 1(F1), and F3 is the field which contains most technologies among these three fields.

Since the nodes can be labeled as different scales, that means nodes from one group can be organized by communities. Some of the classical algorithms will be introduced here for a reference.

- Newman's measure of Modularity [15,16,17]

Based on the model of a bipartite network, apart from the relationship between groups, it is also necessary to have some idea about the features between the nodes in one group, such as the closeness of different software applications. Therefore, a good method to dissociate different communities is of great importance. In Newman's work, an intuitive way is given to solve this problem. According to the testing results of such an algorithm, it is useful to various sorts of complex network in real life applications. It can be considered as one of the most popular method to handle such a problem, because it is not difficult to see his name in the papers of this area. A brief introduction and implementation will be given as follow:

It is assumed that the unipartite network will only be divided into two sub networks according to the algorithm, because more clusters can be separated via a similar operation. According to the equality

$$M=\frac{1}{4k}\sum_{ij}(A_{ij}-\frac{o_i o_j}{2k})(s_i s_i+1)=\frac{1}{4k}\sum_{ij}(A_{ij}-\frac{o_i o_j}{2k})s_i s_i$$

M is the modularity of a specific network, m is the edge of it, A stands for the link between two nodes, o is the order of a node and s represents the grouping status of a certain node. In this algorithm, the network is divided into two groups via maximize the value of modularity M with the help of the knowledge of linear algebra such as eigenvalue, eigenvector, Cartesian product and so on.

- Guimera's method of the computation [8]

The method introduced in the paper of pollination networks to calculate modularity is based on an algorithm from R. Guimera. Generally, the value of modularity of a certain node implies the extent which it has more links to its own community comparing with the condition that if all the edges of this node is randomly built. Mathematically, it can be represented as:

$$M=\sum_{s=1}^{N}(\frac{I_s}{I}-(\frac{k_s}{2I})^2)$$

where M is the modularity of the network, "N" is the number of modules of the whole network, "Is" is the number of links from other nodes to module s, "I" is the number of all the links in the network, and "ks" indicates the sum of degree of all the nodes in module s. Intuitively, it is a good method to define the relationship between nodes and communities in one of the groups of a bipartite network when the communities are assigned, because it is estimated that this algorithm has an accuracy over 90 percent to identify the module. With the help of the modularity, the potential influence from one node to its neighbours can be objectively demonstrated.

- Lambiotte's special idea about modularity [18]

During the process of bipartite network, when we have a new projection network in our hand, the modularity being figured out is not the only information that is useful, some other methods dealt with the network can also reveal something valuable. Lambiotte's work can be a good example, rather than process all the data in a pure mathematical way, Lambiotte focuses on something more graphical, to be more specific, he emphasis on the number of triangles in the network, which could also demonstrates the relation between nodes.

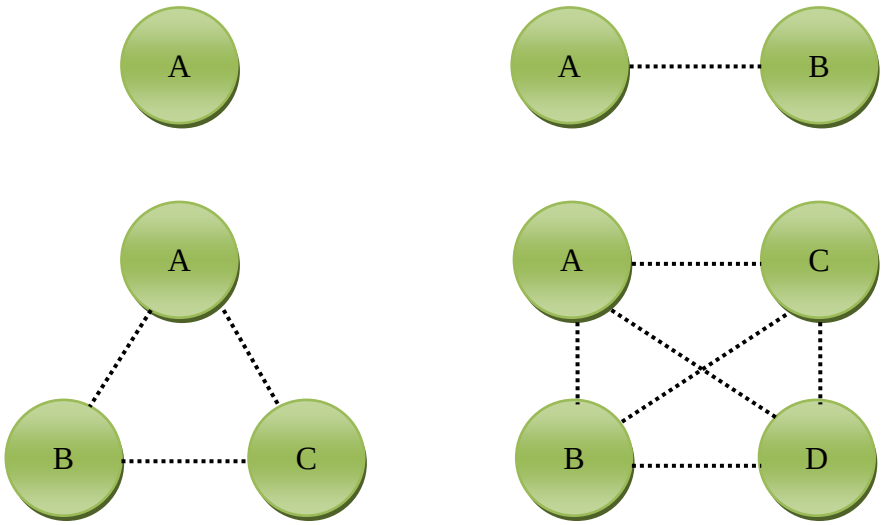


Figure 17. The most basic shapes of the co-operate relationships, namely, from 1 to 4.

Via focusing the basic shapes in a complex network, it would be easier to pick out the sub-communities that the nodes in it are closer to each other than to the nodes in other clusters, thus, once some node is influenced by the external changes, it is easy to draw some conclusion based on the shape discrimination.

In the practical aspect, the usage of modularity is obvious. If the system makes a prediction of a certain node in one group, it should take care of the node that is closely related to the selected one. For instance, if a bug is found in Windows 7, then it is probable that the same bug can be found in Vista and XP, because the operation systems could have something in common. This is the reason why modularity is helpful.

- Multi-level dynamic network

In order to make the model more compatible, it needs to have the ability to deal with datasets of different feature and scale. Feature list is still suitable for all sorts of features, the selected features just have to be time-related to make the result more objective. As for the scale, multi-level network can be a recommended solution. Imagine a dataset with a group of millions nodes inside, intuitively it exceeds the reasonable scale range of the original dynamic network. Therefore, the whole dataset needs to be preprocessed. Generally, with the help of the community computing algorithms, a group with millions of nodes can be probably reconstructed into a 3-level network tree. Each parent node is composed of hundreds of child node, which satisfies the requirement of the scale. On the other hand, it is of great significance to set a value changing threshold between parent node and all its child nodes, in other words, to what extent will the frequent changing of the child nodes influence the value of their parent node is a key point of a multi-level network.

Apart from the features mentioned above, Perez's theory seems to be more compatible with this new model. Specifically, Perez's theory can be added into the matrix both on

a micro view and a macro view. Focusing on a link between two certain nodes, the weight of edge might increase as the shape of the curve mentioned in the theory, although the time period will be influenced by the difference of the features of different nodes. Besides, rise to a higher level, concerntrating on a certain node, despite the fact that it might connect to several nodes from the other group, which the raising speed may probably varies between them, however, the integral raising tendency of the certain node will also majorly follow the rhythm of Perez's theory. On the whole, both of the weight of edges and nodes will change in a similar step, and the node will be influenced by the changing of its own link.

- Dynamic feature list

In the given samples, once the figures in the feature list are set, they will become fixed. It seems to be suitable to some cases, but it is a better practice to make it more general. It is reasonable to set the parameters with the past data, but since the model is created for prediction, it is more convincing for some of the parameters to be changed with the status matrix.

In addition, for some of the dataset, techniques like crawler are necessary to make the figures in the feature list more objective. For instance, in the field of social network service like Facebook, if the feature list is about the type of the sharing route of a hot post at the moment, hence, the crawlers need to collect all the data of the certain post from all the potential pages from the whole platform. From the sight of the crawler, although there is no need to create thousands of crawlers for data collection, but still, it is a big strategy for the whole crawler system.

Specifically, there are several main aspects for the construction of the whole crawler system, such a system is composed of some techniques from the other areas of computer science, which can be introduced as follow. First of all, there has to be a difference of the crawler strategy under different situation, namely, it is reasonable for the crawlers to collect data freely from the platform in normal times, however, when some posts are regarded as a hot issues, it is necessary to gather some of the crawlers to handle the same mission for a high efficiency of the data collection. Secondly, in order to get some data that might be useful for the special needs, hence, the word and image information might be re-constructed and re-stored because some of them might be preprocessed for the calculation of the figures for the feature matrix.

For another, information similarity recognition is important for the whole system, the information of the entire social network platform is exploding every second because of the amount of users, but on the other hand, it is a fact that, quite an amount of information on the web pages are similar or identical. Therefore, it is of great importance to recognize the similarity of different data and store only one version for the information, otherwise, there will be huge amount of redundant information occupying the space of hard disk. On the whole, when data are coming from the source which have a big scale and with a fast updating speed, it is of great

significance to lay down a complete strategy for a better efficiency performance and accuracy of needed figures.

- Chain reaction [19]

According to the purpose of the model, the measure of a single node which will appear or disappear is not comprehensive to some extent. To be more specific, it is also valuable to simulate a dynamic trend based on the current data and prediction results as well. There exists some paper which demonstrate the tendency catastrophic cascade even there is only something seems to be negligible occurs at the moment. Buldyrev demonstrates a special example of the network in his paper.

In terms of bipartite network, an example of blackout is cited for improving. The network of electricity and the network of internet interact and influence each other, namely, the power provides the energy for the internet which controls the power stations. Thus, once a power station is down, it will lead to a repeated process of a cascade of failures. To a more abstract level, it is a similar story for failure delivery.

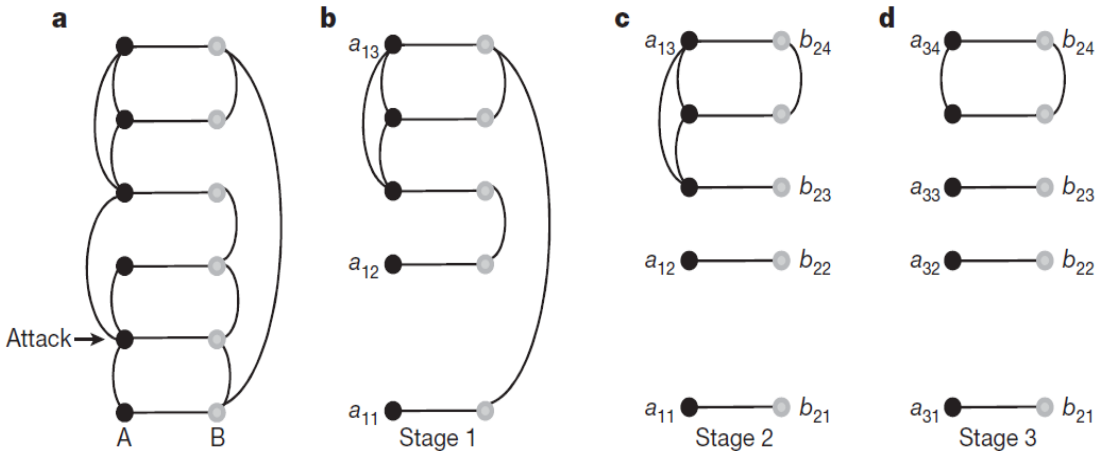


Figure 18. Intuitively speaking, when a node disappears in group A, its links to the neighbors will be removed thus form some new sub-clusters (a11, a12, a13). Relatively, nodes in group B will remove the edges in it in order to make its sub-clusters similar with A(b21,b22,b23,b24, since there is no origin links between b23 and b24, thus four sub-clusters occurs after this step). Group A will give its corresponding feedback again. Such an interaction will keep moving until it becomes stable. In the new project, nodes between groups will not disappear in such a pace, but similar situation should be taken into consideration to make the model closer to real life. Figure 18 is cited from [19].

Another similar example can be given based on the topology of the global macroeconomic network from Lee's work in 2011 [9], the impact of the spreading of economic crises can be evaluated. Generally speaking, the main idea of this crisis spreading model is comparing the weight of edges ΔW lost by a certain node with a set portion (t) of the node's own weight, thus, defined as t*C. The crisis will influence

the node if ΔW is greater than tC. Obviously, research in the field of economic is always easier to arouse others attention, therefore, it is important to choose a field that is close to people's daily life, or at least, being useful to a specific group of people.

- Two direction value changing

Up to now, every concept of the dynamic bipartite network model is based on the assumption that once the edge between two nodes is constructed, it will not be deleted until one of the nodes disappears. From another view, it is necessary to have an assumption about the situation that values in the matrix can be changed from both of the direction, in other words, even if both of the nodes are still in their life span, the link might be cancelled under certain conditions.

Based on the way of timing setting concepts mentioned above, two relative methods will be introduced. From the angle of feature list, apparently, some parameter for declining evaluation need to be added into the feature list. For instance, from the dataset of camera and the techniques on it, if the manufacturer is aggressive on technique updating, there is comparatively a high probability to see the product give up some main trend techniques and try something new or special which does not catch the main focus at the moment. Some other factors like cost, compatibility and competition will also influence the disappearance of a certain link. In addition, find an objective and reasonable data source is also of great importance.

As for the curve from Perez's theory, it is comparatively easier to understand, since the life span of a specific technology on various applications might be quite different. Thus, situations like a certain link of between two nodes has gone to the end of Perez's curve while the development of the node is still on the way of fast developing also seems to be probable. Under such a circumstance, a reasonable solution will be add a time evaluation component on each link of the matrix, hence, when a certain edge is running out of date according to the simulation, the link will be cancelled.

- Data collection method

Under the circumstance which figures of the matrix or feature list need to be calculated from a big scale of dataset, it is not always a relaxing work to collect all the needed data especially some of them are not available on the Internet, or some others have not even been collected before. In order to obtain some special figure, apart from the crawler strategy which has been mentioned above, Foursquare also provide us a good method for data collection. "Keep up with friends, discover what's nearby, save money and unlock rewards", with such a slogan, it is not difficult to conclude this way of data collecting. This is a win-win mode for data collecting, users collect data manually for Foursquare and get discount when purchasing, merchants will attracts more customers because of the discount and the missions from the game-like application from the Foursquare as well, besides, merchants pay for the useful data they want which is collected by the customers, Foursquare creates a platform which provides a data source which is established by the users, they will win the profit from

those who need the processed data collection. It is an effective solution for such kind of problem, but it is not easy to simulate others to get the market, it is a kind of challenge which is from another subject, hence, it will not be introduced specifically.

6.3 Cross field applications

The dynamic bipartite network is originally used for some prediction, in this section, the potential expanding applications will be discussed.

- Artificial intelligence

This can be regarded as a kind of application transfer of the dynamic bipartite network. Since such a network model aims at future prediction, therefore, it is reasonable to apply it on the artificial intelligence system. To raise an intuitive example, such as a strategy game, the computer can simulate the future situation for its own strategy determination. Besides, from the angle of a game, the strength of the computer can be set based on different thresholds and figures in the feature list. As for the data feature and storage, it depends on the specific situation of the game. For instance, for the strategy games related to foreign affairs between countries, since there are many special values that are related to the country's development, and the computer needs to know the current situation of its competitors, it is obvious to find a matrix for development simulation. Overall, artificial intelligence is a direct transfer for the new model that can be easily imagined.

- Image processing

Different with the application in the field of artificial intelligent, image processing can be considered as a kind of application transfer that is field crossing. Specifically, based on the concept of matrix, it is probable to refine a comparatively obscure image to a clearer version, in other words, such an idea tries to raise the resolution of the image without declining its quality. In order to provide an intelligible introduction of this idea, examples will be given during the explanation.

From the view of the digital pictures, it can also be regarded as a kind of matrix with the information of its color. Therefore, based on the theory of dynamic bipartite network, it is possible to do something special on the image. On the other side, differences between figure data and image are also evident, hence, they will be listed at the very beginning. First of all, the two group of an image based on the idea of bipartite network is fixed, namely, X axis and Y axis, every node from a group stands for a position of the image, furthermore, every link between nodes from two groups of the origin image is constructed, it seems to be more like a weighted matrix because the information of color differs between edges.

Based on such an assumption, "similarity" can be used to find the difference extent between pixel lines, specifically, it is reasonable to add a new line of transition color if the similarity between lines is high. Otherwise, it might be a kind of edge of the content in the picture. As for the calculation of contribution, it can be used as a space

optimisation algorithm. When a certain row has a high contribution, the dots of the row are comparatively well ordered by the difference of color. One more thing should note is that under such an assumption, color should be separated into several groups as a preprocessing of the calculation of contribution. Intuitively, the more groups the colors are divided, the higher quality the image will have. Such a practice can be used as an option function after the resolution has been enhanced.

The following pictures will provide a general idea about the procedure of the image processing based on the concept from the dynamic bipartite network, it is allowable to both regarded them as either an image or a matrix with numbers.

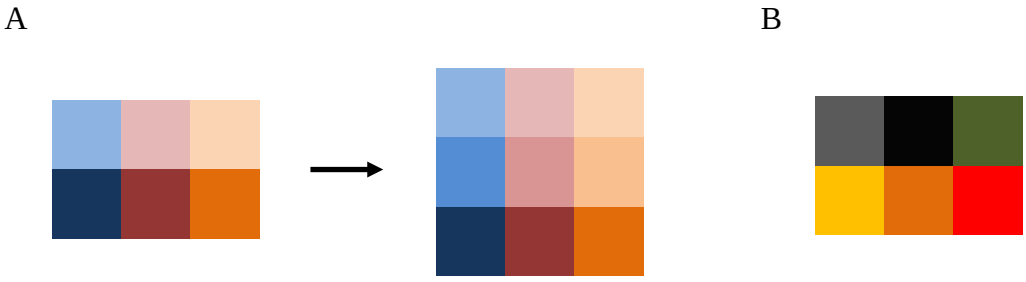


Figure 19. Basic ideas of the similarity between rows. When the similarity between rows is high, such like group A from the figure, a transition line will be given to make the whole image smoother. While the similarity between rows in group B seems to be comparatively lower, there is no need to practice a transition. In addition, it is also a necessity to set some threshold for the judgment.

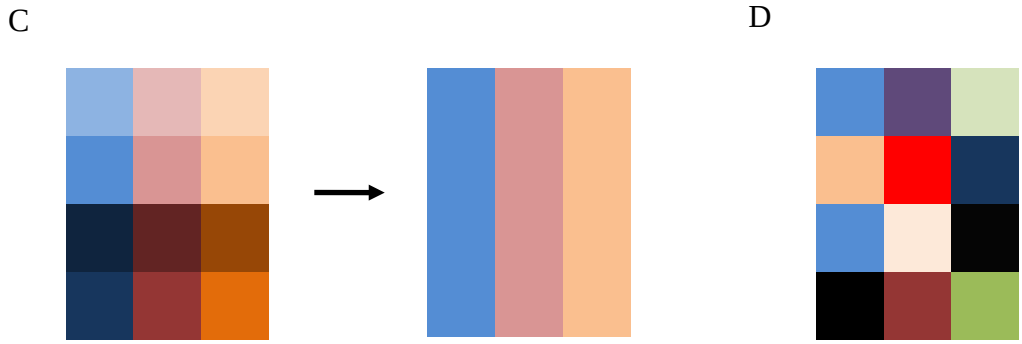


Figure 20. The difference of contribution from the images. If the color are reasonably divided, based on the concept of contribution, since the colors are in the same group, the contribution of the columns in group C will be a high value, thus they can be contracted when it is needed, and it will not be a big difference after the operation. On a contrast, contribution in group D will be comparatively low, therefore, it will look like another picture if it is contracted.

On the whole, it is an application transfer which has not been further researched, it is better to apply it with a real picture for the evaluation. On the other hand, it can be

summarized that the application of the dynamic bipartite network can be expanded to a higher level over the pollination network.

On the whole, based on the basic design of the dynamic bipartite network model, improvements can be found from all the potential aspects, mainly aiming at the scale expanding of dataset, further upgrading of model design and efficiency enhancement. Besides, cross field applications are also attractive. After all, although the listed improvements seems to be reasonable, they need to be implemented and evaluated via some ground-trusted datasets or convincing results such like reduction of time.

Chapter 7
Evaluation of the work

In this part the whole work will be evaluated objectively from different aspects, which will be briefly divided into the following aspects.

7.1 Modeling innovation

Dynamic bipartite network originates from the bipartite network which is chiefly used in the field of pollination. Based on the unchangeable dataset, some special values will be calculated to find the potential species that might be in danger in order to protect them from extinction in reality. Intuitively, this is a process that can be regarded as getting the original data and feedback with some information that is useful. In other words, the origin data is not been modified. Therefore, a new view is displayed here based on the changing of the original dataset. From a general sight, it expands the way of the application of the bipartite network.

In the angle of bipartite network, disappearance is used to be a central issue being focused on. Normally, the disappearance situation of the whole network will be treated only once. While in the dynamic bipartite network, both of the appearance and disappearance will be processed repeatedly in order to excavate the potential usability of the current data to a higher level. At the same time, with the re-understanding of the feature of the dataset and the new model structure, the dynamic bipartite network model is possible to be applied into more fields and areas.

In addition, the time estimating component in the network model can be regarded as the core innovation, such a special design enables the bipartite network upgrades from 2-Dimension to 3-Dimension. From the angle of model structure, assistant matrix for the original dataset is also a kind of innovation. With the help of these characteristic figures, bipartite network is able to provide us a new view.

7.2 Implementation difficulty

With respect of implementation, it seems that the total cost of the model construction is low, even when the new algorithm which is more efficient is designed, as long as the algorithm is well capsulated, the only thing need to do is a substitution. On the other hand, the acceptable performance is based on a small scale dataset, hence, if the model is expected to be compatible with dataset of larger scale, it will cost some time and effort to apply the model into a new system framework.

In addition, it is more important to refine the quality of the linking matrix and the feature list from the original dataset. The more objective and valuable the matrix is,

the more convincing the prediction result will be. From the angle of the user, it is easier to create a relationship network with the data available. On the contrast, it is comparatively hard to choose the features of the feature list and configure them with specific values.

7.3 Accuracy performance

According to the result analysis based on the testing data, such a conclusion can be made that such a design can reach the target of network upgrading. Proper mechanism and threshold values are important to the accuracy of the prediction. Since the model is established with some useful controlling methods, hence the result of prediction is fine. With the limitation of data source and computing ability, the situation of dealing with a larger dataset still need to be proved, with respect to dataset in a small scale, the model can predict the elimination sequence and time gap with a high accuracy.

7.4 Evolution probability

This is another highlight of the dynamic bipartite network, because its evolution probability is high, from the compatible of different dataset scale to the enhancement of the computing efficiency, they are several solutions for the various situations. From an abstract level, improvements can also be separated into model designing improvements and components operation improvements, in other words, both of the theoretical and practical aspects are all on the way for a better performance. Besides, the new designed model can also be used to the other area of computer science, which expands the application coverage of the basic concept of relationship network to a higher extent.

7.5 Summary of prediction models

It is always exciting to find some new predictor. Although it can be regarded as a law that there is not a predictor that can provide a high accuracy prediction in every field, as long as the new model could find the field that it can have a better performance than in other fields, it will become a good news to that certain area because one more assistant tool is found.

On the other hand, the world of predictor is always facing to another problem: iterated loop. This is a latent phenomenon that will lead to an endless loop, in other words, when the user receive a prediction, in order to have a larger profit, they will try to adjust their strategy which will lead to a different result in reality. On the other hand, if the model is able to predict the relative future actions, it will re-process the data and give a new potential result which is closer to the fact. Under such a circumstance, if both of the sides will provide a new result based on the latest action from the other side, the whole system will drop into an infinite loop.

Therefore, such an assumption raises another demand for the selection of the dataset, namely, the user’s action based on the prediction should not influence back to the result. For instance, if there is a relationship between consumer‘s purchasing desire

and product price, intuitively, different price will lead to different purchasing desire, hence, when the consumer is interested in the price prediction, it will lead to a loop between price and marketing scale.

7.6 Conclusion

As a total conclusion of this work, this is a new network model based on the notions of the bipartite network that is designed and proved useful under a certain condition. From the angle of technique development, it still needs to be improved with more practical refining operations and testing of more datasets from different fields and scales. A predictor with a high accuracy, better efficiency and a larger compatibility will be the goal of this model.

Acknowledgements

I am grateful for all the support I received from Dr. Steve Gregory. Sincerely thank for his guidance, understanding and patience during the whole process of writing the dissertation and completing the program. His recommendation of the existing papers enables me to expand my view to have a better understanding of the project. Besides, he reviewed my manuscripts despite of his busy schedule, and his several important suggestions help me to enhance the quality of this paper.

Specially, during the period of selecting the project a couple of months ago, since Dr. Gregory's projects are very popular, thus there were not enough projects to satisfy everyone initially. Under such a situation, Dr. Gregory provided me a new project which needs creativity that stimulates my interest. It is a fascinating project that allows me to add anything possible and reasonable freely into the new model. Therefore, I heartily appreciate Dr. Gregory for creating a nice project theme according to my characteristic.

Finally, this dissertation would not have been possible without all the supports and encouragements from my parents, friends and colleagues. Although the project seems to be a little bit intensive, I am always trying my best to make the project better since people around me show their confidence on me. In addition, some of my friends give me some professional suggestions on the writing of the dissertation, especially thanks indeed.

Bibliography

[1] Tao Zhou et al, Bipartite network projection and personal recommendation, PHYSICAL REVIEW, E 76, 046115, 2007
[2] Jianxi Gao et al, Robustness of a Network of Networks, PHYSICAL REVIEW LETTERS, PRL 107, 195701 (2011)
[3] Peng Zhang, Clustering coefficient and community structure of bipartite networks, Physica A 387 (2008) 6869-6875
[4] Werner Ulrich, Nestedness analysis as a tool to identify ecological gradients, Ecological Questions 11/2009: 27-34
[5] Werner Ulrich et al, A consumer's guide to nestedness analysis, Oikos 118: 3-17, 2009, DOI: 10.1111/j.1600-0706.2008.17053.x
[6] Deok-Sun Lee et al, Scaling of Nestedness in Complex Networks, adjaarXiv: 1110.2825v2 [physics.soc-ph] 12 Mar 2012
[7] Miguel A. Rodrıguez-Girones et al, A new algorithm to calculate the nestedness temperature of presence–absence matrices, Journal of Biogeography (J. Biogeogr.) (2006) 33, 924–935
[8] Jens M. Olesen et al, The modularity of pollination networks, PNAS , December 11, 2007, vol. 104, no. 50, 19891–19896
[9] Kyu-Min Lee et al, Impact of the Topology of Global Macroeconomic Network on the Spreading of Economic Crises, PLoS ONE(www.plosone.org) , e18443, Issue 3, vol.6, March 2011
[10] Serguei Saavedra et al, Strong contributors to network persistence are the most vulnerable to extinction, Nature, 13 Oct 2011, vol.478, 233-236
[11] Carlota Perez, Technological revolutions and techno-economic paradigms, Cambridge Journal of Economics 2010, 34, 185-202
[12] Moore's law, Wikipedia, Available: http://en.wikipedia.org/wiki/Moore%27s_law
[13] Thomas H.Coremen, Advanced data structure, Van Edme tree, Introduction to algorithm(third edition), 531-560, 2009
[14] Sebastian Bustos et al, The Conservation of Nestedness Predicts the Evolution of Industrial Ecosystems, arXiv:1203.3796v1 [physics.soc-ph], 16 Mar 2012
[15] M. E. J. Newman, Modularity and community structure in networks, PNAS, June 6, 2006, vol. 103, no. 23, 8577–8582
[16] M. E. J. Newman, Finding community structure in networks using the eigenvectors of matrices, PHYSICAL REVIEW, E 74, 036104, 2006
[17] Michael J. Barber, Searching for Communities in Bipartite Networks, arXiv:0803.2854v1**,** [physics.soc-ph], DOI: 10.1063/1.2956795
[18] R. Lambiotte et al, N-body decomposition of bipartite author networks, PHYSICAL REVIEW, E 72, 066117, 2005
[19] Sergey V. Buldyrev et al, Catastrophic cascade of failures in interdependent networks, nature08932, doi:10.1038, vol.464, 15 April 2010

[20] Yaron Singer, Dynamic Measure of Network Robustness, Electrical and Electronics Engineers in Israel, 2006 IEEE, 366-370, Nov. 2006
[21] Santo Fortunato, Community detection in graphs, Physics Reports 486 75_174, 2010
[22] Ravindra K. Ahuja et al, Improved Algorithms for Bipartite Network Flow, SIAM J. Comput., Volume 23, Issue 5, 906–933, 04 June 1993
[23] Jianguo Liu et al, Degree correlation of bipartite network on personalized recommendation, Int. J. Mod. Phys. C 21, 137, 2010
[24] Fernanda S. Valdovinos et al, Structure and dynamics of pollination networks: the role of alien plants, Oikos, Volume 118, Issue 8, pages 1190–1200, August 2009
[25] Robyn Wethered and Michael J. Lawes, Nestedness of bird assemblages in fragmented Afromontane forest: the effect of plantation forestry in the matrix, Biological Conservation, Volume 123, Issue 1, Pages 125 – 137, May 2005
[26] Michael J. Barber, Modularity and community detection in bipartite networks, PHYSICAL REVIEW E 76, 066102, 2007
[27] Paulo R. Guimarães Jr, Improving the analyses of nestedness for large sets of matrices, Environmental Modelling & Software, Volume 21, Issue 10, Pages 1387-1516, October 2006

Appendices

A1. Name of the nodes in the table 5 and 14

Techniques and fields selected in the matrix can be listed as follow, which has the same sequence as the alphabet order from the table above. Techniques (Column): information security, artificial intelligence, language recognition, computer perception, neural network, hardware system design, distributed computation, compiling system, processor techniques, storage techniques, peripheral devices, word Information processing, image processing and decision support. Fields (Row): software (single machine)- business application, program developing, multimedia, game; software (network related)- business application, game; hardware (single machine)- robot; hardware (network related)- bank card; network platform- SNS (Facebook), E-mail, online shopping(Amazon), information browsing; network service- data saving (netdisk), data searching (Google), cloud service; science research- military, weather, aerospace.

A2. Matrix of second time evaluation

Software\Hardware	CM	CU	CH	RM	RU	RH	DM	DU	DH	GM	GU	GH
Programming platform	1	0	0	1	0	0	1	1	1	0	0	0
Document processing	0	0	0	1	0	0	0	0	0	0	0	0
2D processing	1	0	0	1	1	0	1	1	1	1	0	0
3D processing	1	0	0	1	1	0	1	1	1	1	1	0
Database	1	0	0	1	1	1	1	1	1	0	0	0
Animation processing	1	0	0	1	1	1	1	1	1	1	1	0
Video Processing	1	0	0	1	1	0	1	1	1	1	1	0
Sports Game	1	0	0	1	1	0	1	1	1	1	1	0
Racing Game	1	1	1	1	1	1	1	1	1	1	1	0
FPS Game	1	1	0	1	1	0	1	1	1	1	1	1

Table A2. The table processed from the data in 2006 for the retesting.

A3. Basic design of the programme

The whole program is written in Java, no database like SQL or Oracle is applied, data storage is accomplished through file stream, namely, the dataset is read or written into a txt file. The program is composed of four classes, each of the class is set to deal with different task.

- starter: The entry of the whole program, all the necessary objects are created here.
- listMatrix: This class is constructed to process the majority of the computations related with the bipartite network matrix, including the functions like computing the nestedness of the network matrix, similarity between rows, update the value of a specific position and delete a certain row or column.
- randomColumnCreator: The main function of this class can be understood from its name, it creates all possible columns which have the same number of nodes as the original column. All the possible columns are used to compute the contribution of the specific node.
- mathProcessor: This class is created to deal with some special tasks such as calculating the mean and standard deviation of several numbers, which is also used to assist the process of the contribution of a certain node.
- stepCounter: This class is created to evaluate the potential steps of the connections between two nodes based on the features it obtained from the feature list. There is a sub mathematical model in it to help calculating the result. Intuitively, the design of the sub model will influence the evaluation of the steps.
- nameSearch: This is the class which draws up a unified format of output, which makes the printing results more user-friendly.

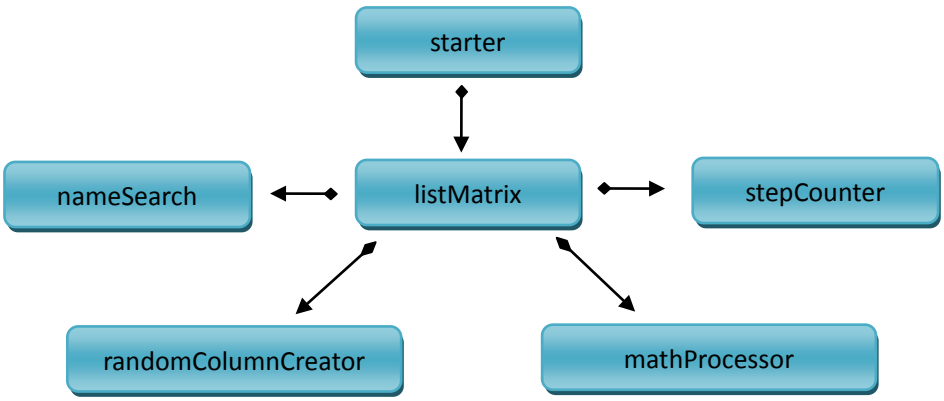


Figure A3. The class diagram of the program. From the class diagram we can see that the status matrix is the core of the model without any doubt. The majority of the computation is related to this matrix, meanwhile, other classes are also important because their existence guarantee the objective of the final result.