Career Analysis by Spatial-Temporal Network
2019 Fall Learning From Data

Group Member:
Yanfeng Chen
Yuan Wang
Shi Mao

Contents

1 Abstract

Career analysis via social connection in both spatial domain (among different people in the same period of time) and temporal domain (among different period of time with regards to the same person) is an interesting and important research area. In this project, we collected government officer resume data, construct a spatial-temporal network representation and analysis the career trajectory based on it. As a result, we make good prediction of institution categories for individuals based on the embedding learned from the structure of spatial-temporal graph.

2 Introduction

Companies such as Liepin, Linkelin and Pinterest have achieve great success nowadays. These companies do help a large group of people to find their suitable jobs. Therefore, improving the methodology behind them becomes researchers' interest.

The conventional methodology utilizes user's resume to extract the characteristics of the career trajectory, such as company, position, education, time, position accumulation and other characteristics of position change, and analyze the influencing factors of position change through historical trajectory data. The data also can be used to train to predict the next position which user would undertake in future.Although it does help connecting people and find jobs with great satisfaction, this is still not the best methodology. It neglects other relationships of the users such as mentors, schoolmates, colleagues and so on.

In this piece of work, we combine the users' career trajectory and their characteristics of their neighbors in social networks to conduct data mining in order to study the possible impact of social relations on the users' career change. More importantly, the application provides users a long term career plan instead of a short period of time.

Recent graph based researches includes generalizing the definition of convolution to non-Euclidean space to directly achieve graph convolution network, learning a good embedding that representing network nodes as low dimensional vectors[1]. It also becomes a very important area of research that how to apply computational analysis of dynamic social networks. Among them, STWalk is a brand new method[2] which can better exploits the spatial and temporal structures to better understand the whole spatial-temporal network.

We believe that the methodology in this research would boost the accuracy of career's trajectory prediction.

3 Data Crawling and Parsing

3.1 Data Crwaling

The data is crawled from the public government websites. For example, in the website http://www.sz.gov.cn/cn/ [3], it contains all of resumes of Shenzhen government staff. With reference to Figure 1, the resume contains the information of Rugui, Chen. It has clearly demonstrated that job transitions of Rugui,Chen from the September of 1979 to now. The locations and institutions of each occupation through these years are also listed.

深圳市委副书记，深圳市人民政府市长、党组书记

陈如桂，男，1962年9月生，汉族，广东廉江人，1992年10月加入中国共产党，1983年8月参加工作，研究生学历，工学博士，高级工程师，享受国务院特殊津贴专家。

1979.09—1983.08 桂林冶金地质学院物探系地球物理勘探专业大学本科学习
1983.08—1986.08 桂林冶金地质学院教师
1986.08—1989.05 中南工业大学地质系应用地球物理专业硕士研究生学习
1989.05—1991.01 广州市建筑科学研究所干部
1991.01—1992.09 广州市建筑科学研究所测试技术研究室副主任
1992.09—1992.10 广州市建筑科学研究所测试技术研究室主任
1992.10—1998.03 广州市建筑科学研究所副所长（其间：1995.09—1998.02中南工业大学地质系应用地球物理专业博士研究生学习）
1998.03—1999.08 广州市建筑科学研究院院长、市建筑集团有限公司副总工程师
1999.08—1999.12 广州市建筑集团有限公司副董事长、副总经理
1999.12—2000.06 广州市建筑集团有限公司党委副书记、副董事长、副总经理
2000.06—2001.07 广州市建筑集团有限公司董事长、总经理、党委副书记
2001.07—2003.05 广州市建设委员会副主任（广州正局级）
2003.05—2007.03 广州市建设委员会主任、工委书记
2007.03—2010.06 广州市人民政府秘书长、党组成员、办公厅党组书记
2010.06—2011.12 广州市委常委、秘书长
2011.12—2015.09 广州市委常委，广州市人民政府副市长
2015.09—2016.09 广州市委副书记、政法委书记
2016.09—2017.01 中山市委书记、市人大常委会主任候选人
2017.01—2017.07 中山市委书记、市人大常委会主任
2017.07—2017.08 深圳市委副书记，市人民政府副市长、代理市长、党组书记
2017.08—今 深圳市委副书记，市人民政府市长、党组书记
广东省第十二届省委委员

Figure 1: resume of Rugui, Chen

3.2 Data Parsing

In order to conduct data mining for analysing the user's career trajectory as mentioned in the motivation part, temporal graph are going to be used later. This will elaborate in detail in next section. Before building temporal network, we need to

extract the key information from the resumes. With reference to Figure 1, the information in the resume is formulated in pattern. It has four main features, which are time duration, occupation location, occupation institution and occupation position. In order to extract these main features in each resume, natural language processing has been used. Natural language processing is a mechanism that helps computer to understand or analysis human language [4]. It has been used in domains such as grammar induction, sentence breaking and parsing [5]. A Github repository contains code for parsing the information. However, some resumes lack of information. After parsing by the using [6], manual adjustment is applied to all the data.

After adjustment, an example of output is in Figure 2. The data has been fully parsed into four parts.

Figure 2: Final Data

4 Spatial-Temporal Graph Construction

To build temporal Graph from collected data, we detect those anomaly collected by mistakes and abandon them to have a clean dataset.

4.1 Connection Construction

Initially, we want to define two person is 'connected' if they work for the same institution in the same period of time considered. However, by looking at the data, we found that of all 720 institutions there are on 108 exact institutions appears more than once, with the most frequently appearing counts 13, including the times that appears in a person's career more repeatedly. Therefore, the connection will be too sparse if we insists on this criteria, resulting in undiscovered relationships.

To address this problem, we turns to measure the 'distance of institution' instead. We first observe that although there are diverse institutions, their name is relatively similar by consisting similar words like 'transportation', 'construction', 'environment' etc. This can be shown clearly by the word cloud below:

Figure 3: the words frequently appears in institution name

Therefore we simply measure the distance by Fuzzy comparing the name of pairwise institutions, and take the matching score as the distance metric. The Fuzzy matching we use here is based on Levenshtein Diance, which basically measures the minimum number of singe-character edited to equalize two strings[7]. Since here we directly use distance, the maximum matching score 100 is mapped to distance 0, while the minimum 0 is mapped to 100 by $dis = 100 - score$. With the measured 'distance', we build the graph using $\epsilon - neighborhood$ by setting a threshold and only accepting the connection whose distance is below the threshold. To test the effect of different connectivity, we build the graph under threshold 20,50,80 separately. Note that we didn't use KNN to build the graph here because inherently, different person tends to have different 'range of social network'. Therefore, it's unreasonable to suppose a fixed 'range' for them.

For each time-step, the constructed graph is a weighted un-directed graph with nodes representing a person with a feature of all his/her job situation, namely institution,

position pairs, and edges representing the minimum institution distance mentioned above. We use minimum by consider that even one may have different job situations at the same time, the connection is established between individuals once they serve the 'similar' institutions. Which means their overall job situations need not to be all similar (measured by 'mean' distance).

4.2 Time-step Selection

The observation of the job starting time distribution is not uniform, which emphasis on recent years approximately from 2012-2019 instead, is because the crawled data from current government official websites usually have more recent resume than far before. Base on this investigation, we choose to use a smaller time interval to investigate the evolution more clearly. Another interesting investigation of dataset is that the birthday of the collected people are around 1960s - 1970s which indicates that they are concentrated on about 50-60 years old now. Indicating that our data is in representative of senior government officers.

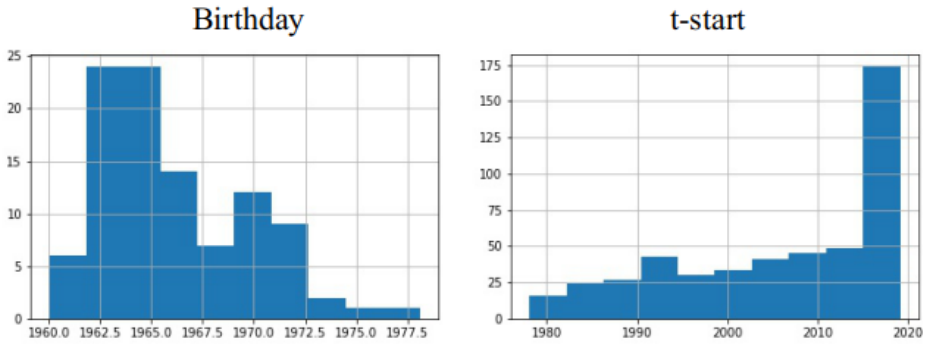

Figure 4: the frequency of Birthday and job starting time

4.3 Labeling

As mentioned above, the diversity of institution makes it hard to make a trajectory prediction for certain person. Inspired by the fact that the constitutional words in institution name is similar, we compromise the prediction task by predicting the 'class' of institutions instead of specific institutions. To create a taxonomy of institutions, we take reference of the Chinese ministerial level department's name, especially those under state council, and categorize all the institutions into these areas.Referencing [8], the complete list of categories are listed in table1, including 54 state council governed departments and three other departments including army,

higher educational institutions and communism party. Note that we separate 'human resource and social welfare' department into two categories because they have distinct range of responsibility.

To efficiently categorize all the institutions into these categories by keyword matching. By looking into the frequently mentioned words in the institution stem, we manually select those words that both appear frequently in the institution names and within the range of responsibility of the corresponding categories department. These words are selected to be exclusive. Then by searching these words in the institutions we are able to categorize them into corresponding class. By doing this we managed to categorized 720 distinct institutions into 58 classes, leaving 100 'hard' institutions like 'commend center' un-categorized.

5 Spatial-Temporal Graph visualization

To have a better understanding of the constructed temporal graph, we provide some visual analysis for both spatial and temporal domain:

5.1 Spatial Domain Visualization

In the sense of data structure, we construct an individual graph for each time-step, representing the people's relationship within this particular period of time. We show the graph constructed by different distance threshold here, where the distance metric is mentioned in section 4.1. We can observe that the density of connection is closely related to the choice of this threshold.

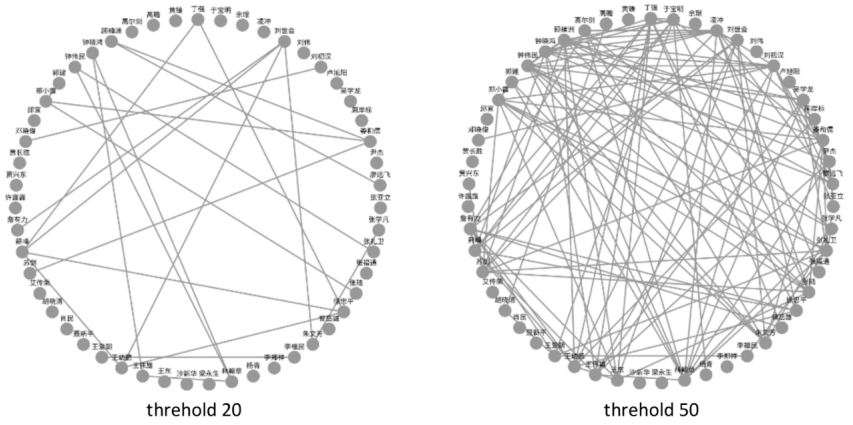

Figure 5: Graph visualization in spatial domain of time-step 1990-1992, with different threshold of 20 and 50

The advantage of having large threshold is clearly that we can exploit more potential relations among nodes. However, since the distance metric is noisy, we can't really rely on this dense connections. Also, by exploring the time domain relations, the drawbacks of sparse connection in threshold 20 is compromised, and its benefit of reliability distinguishs itself. Therefore, we choose threshold 20 as our baseline.

5.2 Temporal Domain Visualization

In temporal domain, we are not actually considering a single graph, instead, we are considering a stack of images. The connection is established by the same person share by different graphs. The temporal domain visualization can show the evolution of a graph. As can be shown below, the connection is evolving if we consider a single node. This allow us to discover the connection 'spatially' but also 'temporally', which serves as a brand new dimension to be analyzed.

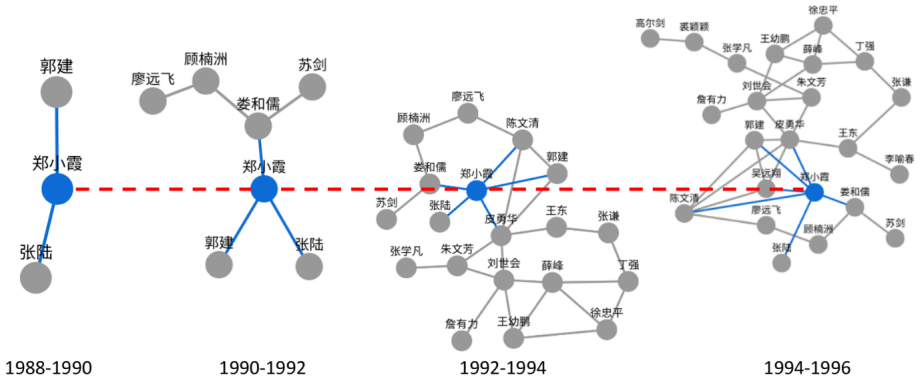

Figure 6: Graph visualization in temporal domain across 1988-1996, by the node 'Xiaoxia Zheng'

6 Data Analysis

6.1 STWalk

In a specified time window, STWalk algorithm can learn a representing vector for each nodes using the structure of the temporal graph. The spatial and the temporal features will be embedded into the vectors in an unsupervised way. The algorithm contains two main parts. One is a random walk on the temporal network, in which way to obtain the spatial and temporal structure. Then the paths generated by the walk procedure will be regarded as sentences where nodes of graphs are regarded as words. We apply SkipGram network to learn the representation of the "words", namely nodes. The SkipGram algorithm can find the best representation for a word to maximize the probability of co-occurrence of two words [9].

To take both spatial and temporal features into consideration, STWalk first construct a Space-Time Graph. The space-time graph is a composed graph for each nodes. To construct such a graph for a node, we need to link the node with itself in the previous graphs to be a series in a fixed time window. Only the first order and the second order neighbors of the node in the previous graphs will be included. In other words, we construct the space-time graph for a node using such ingredients: the current graph and a set of previous subgraphs in the time window. The subgraphs contains the node and its first and second order neighbors. Forexample, if the time window is five, then we will have a current graph and four previous subgraphs. Then we need to link the node in current graph and in previous subgraphs according to the time order into a string. Once the space-time graph is constructed, we can

learn the representation for the node in a new graph, viz. the space-time graph. Thus, the embedding representation for the node will be influenced by both spatial and temporal structures. An example of such graph with window size 4 is shown in Figure 6.

Then we apply a random walk process starting from the given node on the space-time graph to get a sentence-like sequence formed by the visited nodes. When applying the random walk process, the weights of edges will be taken into consideration to make the random walk a weighted random process. In such process, if two nodes share many neighbors or edges, they will be more likely to occur in a sequence together, meaning they should be close to each other in the embedding space. The sequence will be treated as a natural sentence. It is a classical Natural Language Processing problem to embed words from a large number of sentences. After we apply the random walk process, we can obtain a set of node sequence data. So we can apply SkipGram algorithm to learn node embedding representations.

6.2 Classification

Each node represents a real person. The person's class is labeled before. Every node has one or several labels. So it is a typical mutilabel classification problem. First, we need to use one-hot encoding to encode labels into vectors. That means for every kind of label, there exists a dimension. If the node is belong to the class, the value of this dimension will be one. Otherwise, the value will be zero. So the number of dimensions of the vector is exactly same as the number of the label kinds. Thus, each node is corresponding to a vector representing which class or classes the person belongs to.

Then the One-Vs-The-Rest policy is performed. For each label, we construct a classification discriminant to decide whether the node belongs to this label. Thus, the number of classification discriminators will be exactly same as the number of label vector dimensions. For each label, the task turn out to be a binary classification problem.

In particular, we use the STWalk embedding vector for each node as the input data. By the means of leaner Support Vector Machine, we put forward a classification method. The STWalk will embed several ten-year graphs to get an embedded vector for each node as the input data. We extract the label of each node every ten years as the label for the input data. Then seventy percent of the data is used to train the One-Vs-The-Rest leaner Support Vector Machine. The rest of the data is used to test the performance.

7 Result

For the STWalk part, we obtain a set of embedded vector for each node every ten years. In total, we have raw data in forty years. Thus, after the STWalk procedure, we have four datasets. In the data, every node has a corresponding representing vector with sixty-four dimensions. For the label data, after the one-hot encoding procedure, each node has a corresponding thirty dimensions vector. Although there are fifty-four labels in total. The reason is that we only encode the label of the last year of a time window. That means we only gather the labels of nodes every every ten years as the true labels of the nodes. Thus some labels do not appear in this procedure.

The result shows in average, the accuracy of classification is 91.43%. For each label, the accuracy is shown in Figure 7.

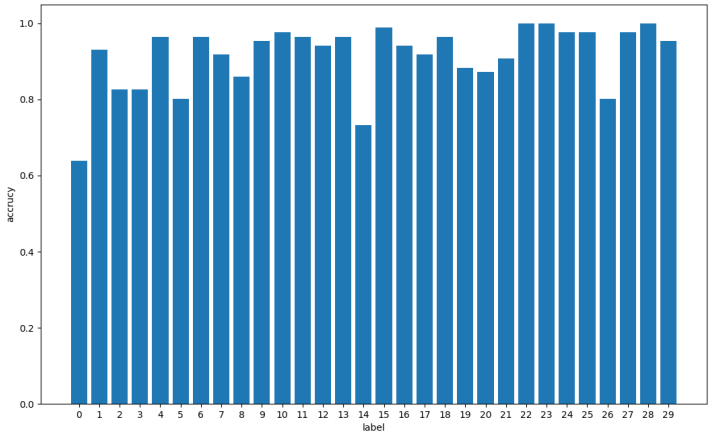

Figure 7: Accuracy of each label

We can see that the accuracy of most labels is high, only the first label's accuracy is lower than seventy percent. This result show our method for classify the node is efficient and accurate in most cases.

We also compute the accuracy of correctly predict at least one label for a node. That is a really difficult task for that how many labels a node has is not definite. The result shows that the accuracy is 52.33%. This result proves that our method has a great ability to classify the nodes using their embedded representation.

8 Conclusion

In this project, we collect a dataset of government officers resumes, construct a spatial-temporal network based on their latent connections, and analysis their career trajectory based on the network. The dataset is crawled form public government website, including the starting and ending time of officers working period, their serving institutions, locations and positions. To construct the spatial-temporal network, we measure the distance between institutions and define the distance between nodes as the minimum institution distance, label the institutions based on the taxonomy of Chinese government institutions and aggregate the collected data by certain time period. We perform an embedding of this spatial-temporal network by constituting a random sentence through random walk in the spatial-temporal network and utilize Word2Vec embedding, which embeds the inner connection in both spatial and temporal domain. Finally, we exam the validation of this embedding by performing a multi-label predicting task, the results shows an average 91% class accuracy and 52% node accuracy.

9 Future Scope

This research only conduct data mining by using resume of Shenzhen senior government staff. Which can be extended to over the country and have more diversity in age. Also, there's another interesting task of promotion prediction, by analyzing the collected information position of their trajectory. This however requires a clear labeling of mapping the institution, position pair to the 27-level civil servant rank. It's beneficial to analyze the speed of promotion with regards to the institution categories and person's social connection. Moreover, although the learning method in this research is a brand new embedding algorithm, it does not take account of any features of the node. In future work, STWALK can be improved in the aspect of exploiting the characteristics of nodes and edges.

A Appendix

A.1 institution Class

根部门	名称
国务院组成部门	中华人民共和国外交部
	中华人民共和国国防部
	中华人民共和国国家发展和改革委员会
	中华人民共和国教育部
	中华人民共和国科学技术部
	中华人民共和国工业和信息化部
	中华人民共和国国家民族事务委员会
	中华人民共和国公安部
	中华人民共和国国家安全部
	中华人民共和国民政部
	中华人民共和国司法部
	中华人民共和国财政部
	中华人民共和国人力资源和社会保障部（人力资源）
	中华人民共和国人力资源和社会保障部（社会保障）
	中华人民共和国自然资源部
	中华人民共和国生态环境部
	中华人民共和国住房和城乡建设部
	中华人民共和国交通运输部
	中华人民共和国水利部
	中华人民共和国农业农村部
	中华人民共和国商务部
	中华人民共和国文化和旅游部
	中华人民共和国国家卫生健康委员会
	中华人民共和国退役军人事务部
	中华人民共和国应急管理部
	中国人民银行
国务院直属特设机构	中华人民共和国审计署
	国务院国有资产监督管理委员会
	中华人民共和国海关总署
	国家税务总局
	国家市场监督管理总局
	国家广播电视总局
	国家体育总局
	国家统计局（副部级）
	国家国际发展合作署（副部级）
	国家医疗保障局（副部级）
	国务院参事室（副部级）
国务院部委管理的国家局	国家机关事务管理局（副部级）
	国家信访局（副部级）
	国家粮食和物资储备局（副部级）
	国家能源局（副部级）
	国家国防科技工业局（副部级）
	国家烟草专卖局（副部级）
	国家移民管理局（副部级）
	国家林业和草原局（副部级）
	国家铁路局（副部级）
	中国民用航空局（副部级）
	国家邮政局（副部级）
	国家文物局（副部级）
	国家中医药管理局（副部级）
	国家煤矿安全监察局（副部级）
	国家外汇管理局（副部级）
	国家药品监督管理局（副部级）
	国家知识产权局（副部级）
高等院校	高等院校
中国共产党	中国共产党
行政机关	机关

Table 1: Chinese ministry level departments

机构类别	关键词
外交	外交
国防	国防
发展和改革委员会	发改委 发展 改革
教育	语言文字 教科文 考试院
科学技术	科技 科学 技术 高新 专家
工业和信息化	工信 工业 信息化
民族家族事物	民族
公安	派出所
国家安全	国安
民政	民政
司法	司法 公证 律师 监狱 监察 检查 法规 法律 政法
财政	财政 经济 投资 国库 预算 会计 经贸 工贸 贸工
人力资源	人事 人力 组织部
社会保障	社保 慈善 社会事务 救助 救护 工会 保险 福利 捐助
自然资源	海洋
生态环境	环境 生态环境 人居 环保 环境保护 污染
住房	建设 住房 建筑 房
交通运输	交通 运输 港务 港航
水利	水务 水源 河湖 水文 水质
农业农村	农业 农村
商务	商务
文化	旅游 文化旅游
卫生健康	卫健 卫生 健康
退役军人	退役军人
消防	应急管理 救援 消防 三防
银行	银行
审计	审计
国有资产监督管理	国有资产 国土监管 规划 国土 城市设计 国资
海关	海关
税务	税务
市场监督	监督管理 监督 市场 监管
广播电视	广电
体育	体育
统计	统计
国际发展合作	发展合作
医疗保障	医保
参事	参事
机关事务管理	机关事务管理
信访	信访
粮食物资储备	物资 粮食 储备
能源	能源
国防科技	国防
烟草	烟草
移民	移民
林业和草原	林业 草原
铁路	铁路
航空	航空
邮政	邮政
文物	文物
医药管理	医药管理
煤矿监察	煤矿监察
外汇管理	外汇管理
药品监督	药监 医药监督
知识产权	知识产权
部队	军 警备
高等院校	大学 学院 学校
党	常委 政治 共青团 市委 区委 团委 县委 省委 委员会
机关	行政 管理 街道 政府

Table 2: The institution categories and their related keywords

References

[1] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S Yu. A comprehensive survey on graph neural networks. *arXiv preprint arXiv:1901.00596*, 2019.

[2] Supriya Pandhre, Himangi Mittal, Manish Gupta, and Vineeth N Balasubramanian. Stwalk: learning trajectory representations in temporal graphs. In *Proceedings of the ACM India Joint International Conference on Data Science and Management of Data*, pages 210–219. ACM, 2018.

[3] resume. `http://www.sz.gov.cn/cn/xxgk/zfxxgj/sldzc/sz_97404/crg/`. Accessed: 2019-12-30.

[4] James F Allen. Natural language processing. 2003.

[5] nlp. `https://en.wikipedia.org/wiki/Natural_language_processing`. Accessed: 2019-12-30.

[6] Han He. HanLP: Han Language Processing, 2014.

[7] Fuzzy matching tools: Fuzzywuzzy. `https://github.com/seatgeek/fuzzywuzzy`. Accessed: 2019-12-30.

[8] Chinese organizations. `http://www.scopsr.gov.cn`. Accessed: 2019-12-30.

[9] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.