Visual Comparison for Department Clusters

Shidi Yu, Amy Woods

December 2020

I. Introduction and Motivation

Network data has become exceedingly important nowadays as it can provide valuable information and insights in a wide variety of fields ranging from bioinformatics, to text analysis, to social sciences. There has been a significant amount of research done on how to create visual analytical tools that can help a user analyze a single network. However, methods of comparing multiple networks, and effective methods of visualizing similarities and differences are two topics in need of thorough research.

One of the first aspects of this research that needs to be addressed is how to quantifiably measure the similarities and differences between two networks. This is a complex problem in itself as there have been numerous approaches to a solution. The methods implemented range from simple graph theory statistics to machine learning algorithms. The difficulty that arises from these solutions is determining which attributes or characteristics of a network are considered essential when comparing two networks. The answer becomes convoluted as increasingly multifaceted and versatile networks appear. Even if a standard was created to specify which attributes are proven to be most significant in network comparisons, such as statistical measures, centralities, graphlets or a combination of these, it still leaves the problem on how to meaningfully visualize such comparisons [1]. Numerous algorithms have been created to compare the networks, but little work has been done to visualize the similarities and differences.

As a result, we created an interactive visual analysis system for not only analyzing the internal components of a single network, but also for comparing networks. Specifically, the dataset that we selected is the Facebook data of students from Taiwan Universities. Firstly, there is a survey data, which includes pertinent information (e.g. school, department, discipline, and gender) of the Facebook account owners. We denote each owner as EGO, and the other owners related to this owner are ALTERs. These EGOs and ALTERs are the nodes in the constructed network. Secondly, our dataset includes the interaction information of two owners. Each line of the interaction data first includes the source and the target of this interaction, for example owner1 (source) commented owner2 (target). Then it also includes the type of interaction (such as like, message, and photo tagging), and the timestamp of this interaction. We used the types of departments to initially cluster the data into a subdivision of networks. In comparing and analyzing these networks, we had the goal of using the system to gain insights on what network departments are popular, and what trends determine that a department network is popular.

II. Related Works

The key tech we used for this project is Contrastive Network Representation Learning (cNRL). There are three main parts included in this project: Graph Embedding, Contrastive Learning and visual guidance for comparison.

Graph embedding, or network representation learning (NRL), is a technique to learn low-dimensional vectors to represent a complicated network. Network nodes or links usually have multiple attributes (e.g. degree, betweeness, closeness and pageRank). It's not easy to find nodes that are similar in terms of structure and attribute in the original high-dimensional network. But when networks are projected as 2-D or 3-D scatters, they are easy to interpret. Some commonly used NRL methods include node2vec, DeepGL and some deep neural networks [1].

Contrastive learning (CL) aims to find the feature that is the most salient in the target network compared to the background network. A network usually consists more than ten features, and reviewing the value of each feature is troublesome and not convenient for users to perceive the difference between networks. With CL, we can get the most important feature of the target network and then find the major difference between the target network and the background network in terms of both structure and attributes. Popular CL methods for representation learning includes contrastive PCA (cPCA), contrastive variational autoencoder (cVAE) and contrasting clusters in PCA (ccPCA) [2].

Visualization for Network Comparison includes three basic types of positioning methods: juxtaposition, superposition and explicit encoding [3]. Juxtaposition is to put two networks side by side and see their differences. Superposition is to position two networks in an overlapping way. Explicit encoding means to directly highlight the different parts of networks.

DeepGl, ccPCA and juxtaposition were used in the system of this project. Specially, we added one more feature to describe the ratio of external interactions. Network representation learning with this feature shall potentially give us information of department popularity.

III. Feature Interpretation in cNRL

DeepGl and ccPCA are the NRL and CL method for cNRL in this project. cNRL starts with DeepGl setting basic features and then learning a multi-layered hierarchical graph representation where each successive layer leverages the output from the previous layer to learn a higher-order feature [4]. In this project, we keep all the basic features and high-order features

to see the structural difference of networks. We also added the external interaction ratio as one of the features to see if the popularity of department is a major factor for the difference.

A. Base features

Degree: The number of direct neighbors of a node.

Betweeness: Betweeness shows the importance of a node as a bridge. If a node is on the shortest path of other node pairs, it acts as a bridge.

Closeness: Closeness is calculated as the reciprocal of the sum of the length of the shortest paths between the node and all other nodes in the graph. It shows how close a node is to other nodes in the network.

EigenVector: EigenVector measures the influence of a node by its relative score. Relative scores are assigned to all nodes in the network based on the concept that connections to high-scoring nodes contribute more to the score of the node in question than equal connections to low-scoring nodes. A high eigenvector score means that a node is connected to many nodes who themselves have high scores.

PageRank: PageRank works by counting the number and quality of incoming edges to a node to determine a rough estimate of how important the node is. The number of incoming edges are counted recursively.

Katz: Katz measures the influence of the owner by calculating the relative degree of the owner within the social network. The relative degree here means to calculate the number of the immediate neighbors (first degree nodes) and also all other nodes in the network that connect to the node under consideration through these immediate neighbors. Connections made with distant neighbors are, however, penalized by an attenuation factor.

B. Higher-order features

For higher-order features, we have operators mean, sum, maxium, Hadamard, WeightLp, and RBF. By applying these features to two-hop or further away neighbors, the higher-order features can be calculated. They are helpful to retain the structural information of the original network.

C. Node-wise attribute: External Interactions Ratio

In an attempt to determine which departments are considered popular, we decided to add a feature to the machine learning algorithm. However, the question remained of what quantitative feature represents the popularity of a department. In our research, we determined the typical definition states that a person is popular if they are known and liked by a large number of people. In terms of popularity of a department network, there are two types of popularity to consider: internal and external. A department could have numerous internal edges and few external edges, or vice versa. Simply counting the edges would not necessarily be an indicator of popularity between separate department networks. For our system, we decided to define departmental network popularity based upon external, rather than internal interactions. The number of different departments the "popular" department is connected to is not currently considered. As a result, the feature we added to the machine learning algorithm is (number of external interactions / total number of interactions).

The representations are then exported as a json file. Each pair of two networks is with a set of representation scatters. For each pair, the pair id and corresponding target, background department id are included. The coordinates, most important feature value, and the external interaction ratio of scatters are also included per pair item.

IV. Implemented Design

A. Cluster View

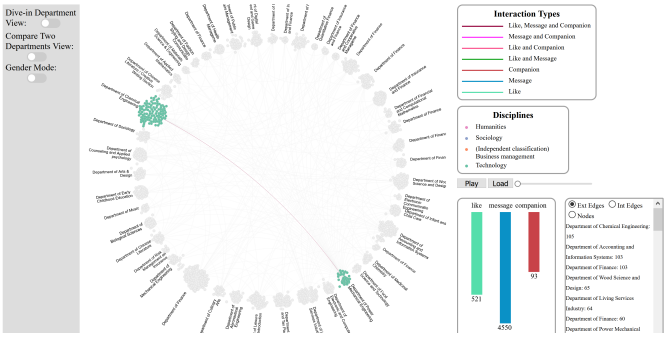

Fig. 1. Cluster View of Dep of CE and Dep of PME

The initial display is a ring of clusters. Each cluster represents a department included in the data. The ring structure allows constant representation of all the departments at once throughout the system. Each cluster is uniquely colored by the discipline they belong to, as shown in the Disciplines Legend. Additionally, in this cluster view, there are lines connecting one department cluster to another. These lines represent the overall interactions between each pair of departments. The lines are color coded to show what type of interactions occur between the departments. Specifically, the colors designate the number of likes, the number of messages, and the number of companions (the number of tagged photos). The user is able to select a department which highlights the interaction lines that connect to the chosen department. Next, the user can obtain a detailed view of interactions between the current department and another department by selecting a second department(Fig. 1). The Detailed Interactions Info Panel displays this. This view allows users to compare the number of interaction types and determine the closeness of two departments. The closeness of two departments can be determined by the logic that companion>message>like. In other words, a department paring with more companions than other interaction types has a closer relationship than a department paring with the same number of likes. The user can reset their selections at any point by simply clicking any blank space within the ring. The cluster area returns back to the Clusters view when neither the Compare View or Dive In View is selected.

B. Dive In View

While other views compare the data at a department level, this view examines a single department internally at an owner

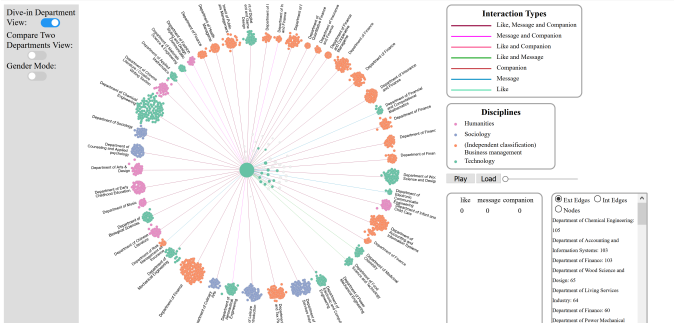

Fig. 2. Dive In View of Department of Power Mechanical Engineering, Large Owner Selected

level. The Dive In view is activated when only the Dive In Toggle is selected. The user selects a department from the ring which expands in the middle of the ring. The circles in the internal view each represent a single Facebook owner and the size of the circle is the degree for that node. The user then has two options to explore this specific department.

Firstly, the user can explore the temporal patterns of the department network by, after selecting a single department, clicking the Load button and then click Play to start the temporal data animation. This animation highlights the nodes that have edges in the current month/year timestamp. The month/year timestamp, the total number of edges, and the total count of interactions are shown near the top of the inner circle. Additionally, the slider can be moved when the animation is paused to show a particular month/year timestamp. This allows the users to see when a department was the most active on Facebook, and who participated in these interactions. This animation can be used to either observe and compare the temporal pattern of the department as whole, or focus on the temporal pattern of a single owner in the department.

Secondly, the user can explore the type and number of interactions at an owner level. Once the user selects a department from the ring, a user can click on any of the nodes from the department to see the internal and external edges. The corresponding nodes that link to these edges are now highlighted (Fig. 2.). Nodes that are highlighted, and appear to have no edges to the selected owner, are Facebook friends of the owner, without having any of the 3 types of interactions. At this point, the user can either select another internal node or an external department to then show the interactions between the two in the Detail Interaction Info Panel. The user can reset their selections at any point by simply clicking any blank space within the ring.

In summary, the Dive-In View allows the user to explore activity of an individual user, as opposed to the department level represented in the Cluster View. The user is able to analyze the closeness of a specific owner to either another internal owner, or another department.

C. Compare View

The Compare View is the key tool for the department comparison task. Like the Dive-In view, two departments are spread out once clicked on and they are Juxtapositioned in the center of the cluster ring. The first clicked department will be the target department, positioned on the left, and the second will be the background department, positioned on the right. Once the network representation scatters for the two departments are loaded successfully, they are shown below the cluster ring. The target department representation scatters are with black border while the background ones are with gray border. On the right side of the representation view is the color legend of the most important feature, feature name, and a toggle to change between the most important feature and external interaction ratio as the color mapping value. On the rightmost side are the detailed statistics panels for the target and background departments. The user can lasso select the representation scatters with left clicking to highlight the corresponding nodes in the original network. In this way, it is convenient to see which nodes make the target network unique compared with the background network. Furthermore, the representation view also support the panning operation with right click and zoom in and out operation with wheel scroll. In this way, users don not need to worry about scatters overlapping and scatters out of border.

D. Listing View

In addition to the 3 main views, there is the Listing View that is linked to the views located in the cluster ring. This view provides a basic ordered list of the departments based on either number of internal edges, number of external edges, or number of nodes. Additionally, when a user hovers over one the listing entries, the corresponding department becomes highlighted in black within the cluster. This view allows the user to not only locate certain departments based on their listing status, but also allows the user to compare the overall structure of all the departments from a less detailed perspective. This view also allows the user to try to detect patterns of departments with certain network characteristics. For example, a user may want to explore if there are similarities between the three departments with the highest number of external links.

E. Gender Mode

The gender mode toggle button changes the color of all the departments, within and on the ring, to have each circle color be representative of its gender. This can provide a quick assessment of overall gender ratio per department. It is also capable of being used to examine possible patterns associated with a certain gender. The Gender Mode button can be turned on at any point to gain different information, whether it is examining a user in the Dive-In view or comparing the gender ratio of two departments in the compare view.

V. Evaluation Results

A. Case Study - Popularity

To evaluate the system, we used the Facebook data described earlier and attempted to see if we could define a popular department in which there was a high number of external interactions. We began with the Cluster View and

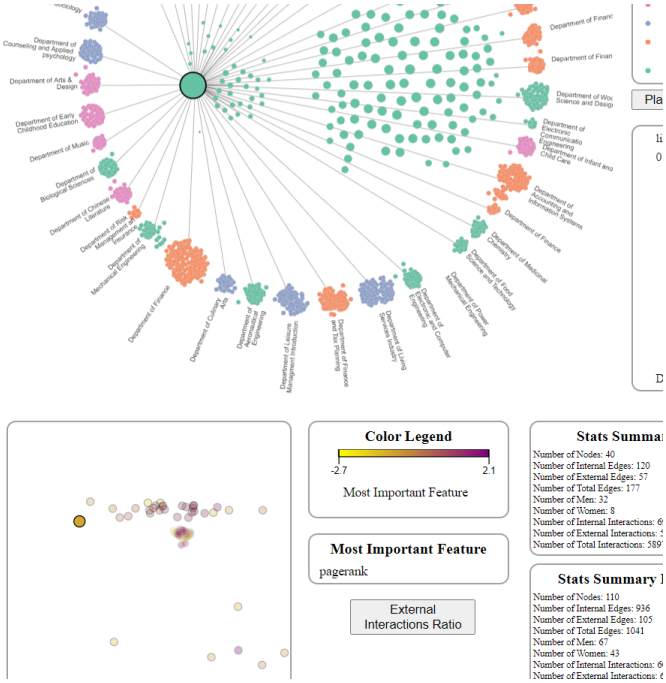

Fig. 3. Low value of PageRank for the big node

Stats Summary Target

Number of Nodes: 40
Number of Internal Edges: 120
Number of External Edges: 57
Number of Total Edges: 177
Number of Men: 32
Number of Women: 8
Number of Internal Interactions: 6902
Number of External Interactions: 52072
Number of Total Interactions: 58974

Stats Summary Background

Number of Nodes: 110
Number of Internal Edges: 936
Number of External Edges: 105
Number of Total Edges: 1041
Number of Men: 67
Number of Women: 43
Number of Internal Interactions: 60699
Number of External Interactions: 6561
Number of Total Interactions: 67260

Fig. 4. Detailed statistics board for Department of Power Mechanical Engineering and Department of Power Mechanical Engineering

Listing View. The information in the Listing View, enabled us to determine the Department of Chemical Engineering (Dep of CE) has the highest number of external edges, and was worth pursuing. We were curious if the more external edges implied more external interactions. When selecting that department in the Cluster View, we noticed the department interacts with several other departments. However, examining the color of the lines led us to the conclusion that these interactions weren't particularly close as many were missing the closest interaction type: companion. One of the red lines, which includes companion interaction, connected the Dep of CE to the Department of Power Mechanical Engineering (Dep of PME). When exploring this red line in the Detail Interaction Info Panel we noticed a significantly higher number of companions compared to other department pairings(Fig. 1.). This implies that these two departments are particularly close in terms of friendship interactions. This led us to explore the Dep of PME in the Cluster View. From this view, we were able to distinguish two unique characteristics about this department. First, it has interactions with all other departments. Second, most of the interaction lines are red which indicates a closeness to many other departments. Therefore, the Dep of PME became our top candidate for possible a "popular department".

Consequently, we wanted to explore the Dep of PME at an owner level and accessed the Dive In View for this department. One of first observations we came across was this department had an unusually large owner node (Fig. 2.). This owner node is large because it has a large degree. We selected this large node and determined that the owner has some internal links, but the internal links were minimal. Additionally, the interactions for these links did not imply closeness, as they were messages or likes, instead of companion. In contrast, the external links of the large owner were numerous and implied closeness to many departments. This indicates that the owner is very active and is well-known among a huge range of departments.

Dep of PME is very interesting to explore, so we chose it together with Dep of CE to compare in the Compare View. Firstly, the major feature is PageRank. From the overview of the original network and some highlighted area with high feature value, we can see Dep of CE is more strongly connected. Secondly, for our purpose of identifying department popularity, there is a very interesting finding. The largest node, or the individual who has the most friends, is with very low value of PageRank (Fig. 3.). As our features are all calculated with the current network isolated (except the external interaction ratio), it means this individual has very low level of interaction within the department, though it has a very high value of external interaction ratio.

Usually departments have apparently less number of external interactions, but it can be observed from our general statistics board that Dep of PME has 52072 external interactions, far more than its internal interactions (Fig. 4.). This large node is the cause of this phenomenon. Probably he or she is very popular among Facebook users.

Overall, based on the analysis, the Department of Power Mechanical Engineering does fit our definition of "popular"

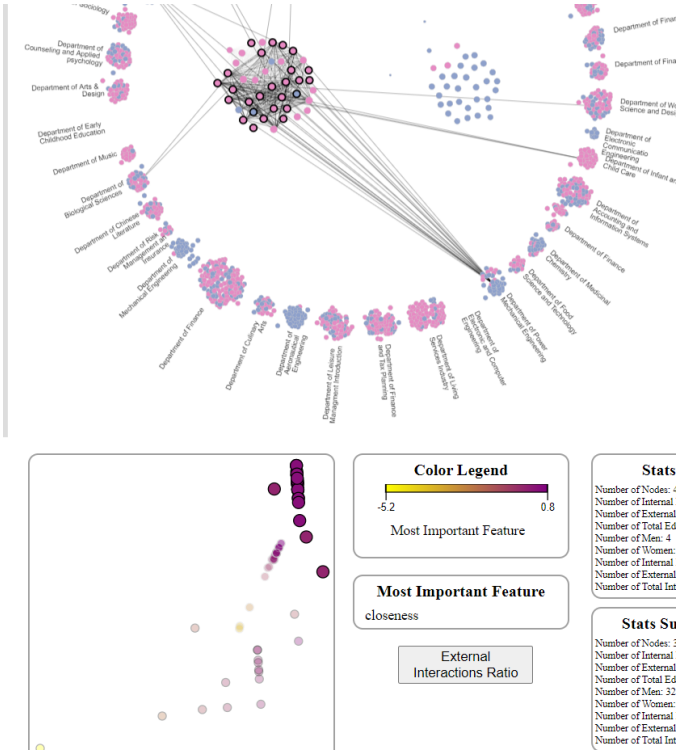

Fig. 5. Closeness difference between Department of Early Childhood Education and Department of Electronic and Computer Engineering

by having many external interactions of close value. This department is unique in structure, too, as we noticed through our Compare View. However, the fact that the department has a single large owner containing a majority of the external interactions makes us consider the department is only "popular" because of the one owner. Our system didn't account for such a huge owner outlier, as we assumed most owner degree sizes would be similar. This questions the validity of categorizing this department as "popular" and deserves further research to determine the actual most popular department.

Another popularity study is to compare Department of Chemical Engineering (Dep of CE) and Department of Financial and Computational Mathematics (Dep of FCM). From the overview of both the original networks and the representations, Dep of CE has a larger amount of external interactions and more nodes with external interactions. In our definition of popularity, Dep of CE is more popular than Dep of FCM.

B. Other Observations - Gender Ratio and Temporal Patterns

Another interesting idea to use our comparison tool is to compare two departments with extremely different gender ratios. Department of Early Childhood Education is a Humanities department with very high ratio of females, while Department of Electronic and Computer Engineering is a Technology department with very high ratio of men.

The result shows that individuals in Department of Early Childhood Education are closer to each other than those in Department of Electronic and Computer Engineering (Fig. 5.), which means the topology of Department of Early Childhood Education is closer to wholly connected. This indicates that females are more likely to add friends on social media.

Another aspect our system can explore is temporal patterns. For example, we determined based on the Temporal Animation in the Dive In View that most departments experience a high number of interactions during the end of 2012 and beginning half of 2013. We are curious as to why most departments follow this pattern.

VI. Lessons Learned/ Further Research

One of the main difficulties we learned about out our technical approach was that the pre-processing part of the feature matrix always generates "standard deviation is close to zero" warning, which makes some representation scatters overlap with each other. The reason of this problem is still unsolved. It could be the uniqueness of our dataset that causes this warning. It should be solved in the future.

We personally found that the Compare View and Dive In View were frequently used together, and it became inconvenient to continually switch between these two views. We learned it would be better to have all three of our main views (Cluster, Dive In, and Compare) viewable at once for convenience of the user. We also discovered that our system does not take into account extreme owner outliers when determining a top popular department. This is an issue that must be addressed as a large owner, such as the one from our case study, can lead to a false belief that a department is popular.

Furthermore, our system only was based on the assumptions that external interactions can determine if a department is popular or not. We believe it is worth considering, analyzing, and adding other attributes to the Contrastive Learning Algorithm, such as number of external edges or the number of internal interactions, to create a more precise definition of a popular department.

In the future we hope that, along with our system limitations being resolved, we will be able to expand the use of the system. Specifically, we would want to enable multiple network comparisons simultaneously, rather than just two networks. Additionally, we would like to cluster similar departments together using techniques such as Exponential Random Graph Models (ERGMs) in statnet. With these enhancements, we aim to provide analysts with powerful methods and tools to compare networks successfully.

References

[1] T. Fujiwara, J. Zhao, F. Chen, and K.-L. Ma, "A visual analytics framework for contrastive network analysis," *arXiv preprint arXiv:2008.00151*, 2020.

[2] T. Fujiwara, O.-H. Kwon, and K.-L. Ma, "Supporting analysis of dimensionality reduction results with contrastive learning," *IEEE transactions on visualization and computer graphics*, vol. 26, no. 1, pp. 45–55, 2019.

[3] M. Gleicher, D. Albers, R. Walker, I. Jusufi, C. D. Hansen, and J. C. Roberts, "Visual comparison for information visualization," *Information Visualization*, vol. 10, no. 4, pp. 289–309, 2011.

[4] R. A. Rossi, R. Zhou, and N. Ahmed, "Deep inductive graph representation learning," *IEEE Transactions on Knowledge and Data Engineering*, 2018.