Extracting Information from the Network Structure of Technology Meetup Groups

Visual Analytics | INM433

Andrew Gibbs-Bravo

180028746

Word Count by Section (excluding headers):

Motivation, Data, and Research Questions	394
Overview of Tasks and Approach	591
Analytical Steps	929
Findings	399
Critical Reflection	958

1 Motivation, Data, and Research Questions

1.1 Motivation

Network structures are ubiquitous and extracting meaningful insights from social network data is crucial for creating a successful social media business. London is one of the world's most active technology hubs and therefore examining the network structures of the technology Meetup groups in London is a very interesting domain to explore what insights that can be drawn.

Meetup is a web service used to facilitate organizing events about a topic and build a community, with over 35 million members as of 2017. [1] Groups range from pickup soccer games to specialized topics such as "Deep Learning with Apache MXNet London".

1.2 Data

In order to understand the relationships between technology Meetup groups, I first created a dataset using the Meetup API by selecting groups which are: located in London, in the technology category, with over 50 members, and that are public. I then created another set of API queries to get the members from each group. The member information contains member id, name, and date joined among other features not used in the analysis.

I created a bipartite graph [2, 3] of the members and their groups and then created a bipartite network projection [4, 5] to get a weighted graph of the shared membership between groups. The shared membership weights are defined as the average proportion of shared ownership in order to account for different group sizes.

Limitations of the Data

As the group data was constructed from current membership data, it does not contain information on members who have left the group or groups which no longer exist. Some groups are "online" groups which means that while they are headquartered in London, they tend not have physical meetups.

1.3 Research Questions

The research questions focus on extracting and applying information from the network graph:

1.3.1 Can you use information from the network graph to determine which members are most likely to be interested in joining a given group?

- Creating a group recommender system would help Meetup grow their platform by increasing user engagement and the framework could be applied by any entity operating within a similar network

1.3.2 Can you use information from the network graph to increase accuracy in predicting membership growth?

- This is important for new meetup groups in terms of both benchmarking growth and managing drivers if they want to continue to grow i.e. where they should be spending their effort in promotion

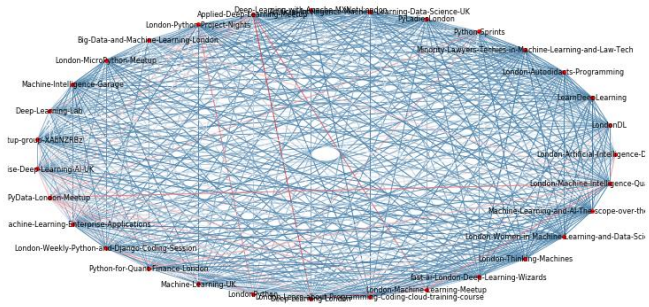
2 Overview of Tasks and Approach

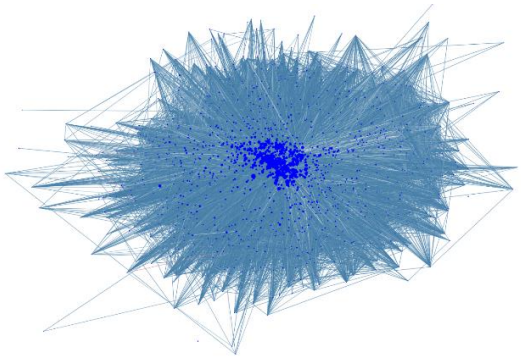
2.1 Can you use information from the network graph to determine which members are most likely to be interested in joining a given group?

2.1.1 Analytical Task: To describe how group structure varies throughout the network and identify groups with similar network structure

The main method consists of developing an interactive network graph to explore the relationship between the different groups inspired by other approaches. [6, 7, 8, 9]

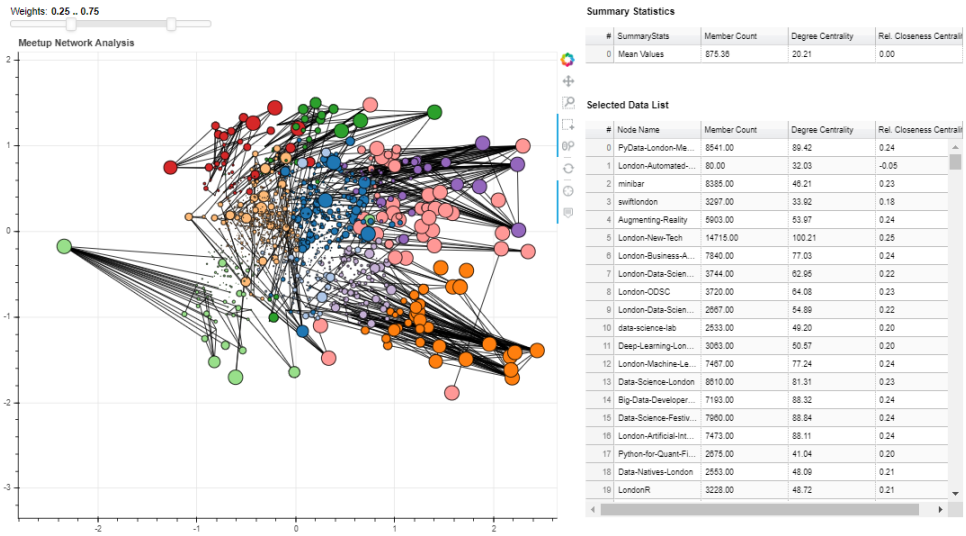
Even with using colors to represent edge strength, simple network graphs were unable to show the information even on a sample of the nodes (below).





Given we are investigating the relationships between groups rather than between clusters, a node clustering approach is not appropriate. [10]

To investigate the relationships more easily, I developed a tool in Bokeh which I named GraphView that contains more information, and more importantly is interactive. The default view is shown below although **please consult the appendix for a full description**.



The visual and computational methods are heavily intertwined in this methodology and inform one another. The initial visualizations were instrumental in determining that the average between proportions was the optimal weighting versus other approaches including not normalizing for group size.

The dimensionality reduction algorithm was selected using a visual inspection of the node positions in GraphView. PCA was also attempted although MDS resulted in more interpretable and intuitive node locations given its ability to maintain distance relationships between nodes. [11, 12]

The node color is based on the assigned k-means cluster which requires the number of clusters as a hyperparameter. The optimal K value is selected based on the interpretability of the clusters visually and the model performance.

2.1.2 Analytical Task: To build models to understand the key characteristics in predicting a given member's meetup group membership and assess the model quality

Achieving a Better Understanding of the Data

To understand the collinearity of the data and for feature reduction we can use a correlation heatmap. [13] Bar charts comparing means can be used for preliminary exploration of the difference between the means of the populations. Pair plots of the clusters and bar charts showing the number of groups in each can be helpful to understand the distributions of the features and interrelations. The proportion of loners in a given group can be explored using another correlation heatmap between the proportion of loners and other variables and using pie charts. [13]

Evaluate the Model Performance

Model performance is somewhat challenging to assess in this use case although visual tools can help analyze model performance. Receiver operating characteristic curves show relative model performance in terms of false positive and true positive rates between algorithms. [14, 15, 16] However, as AUC is not the target metric, a chart showing the probability a member is part of a group versus whether they are in this group would be more useful (named Entropy Waterfall).

The computational model is informed by visual methods in multiple ways including the number of clusters selected through visual and k-means clustering. The selection of models is determined by the Entropy Waterfall in addition to F-measure performance. The feature importance can also be interpreted graphically to understand the importance of each feature. [17, 18]

2.2 Can you use information from the network graph to increase accuracy in predicting membership growth?

2.2.1 Analytical Task: To describe how group membership changes over time and identify groups with similar growth behaviour

Investigating changes in membership lends itself to approaches from the time series domain in which line charts are commonly used to understand this data. [19] In order to visualize the data, I will first build a line chart with all time series data and will apply clustering if this is not visible and then use visual inspection to determine the quality of the clusters in combination with analyzing model performance. [20, 21]

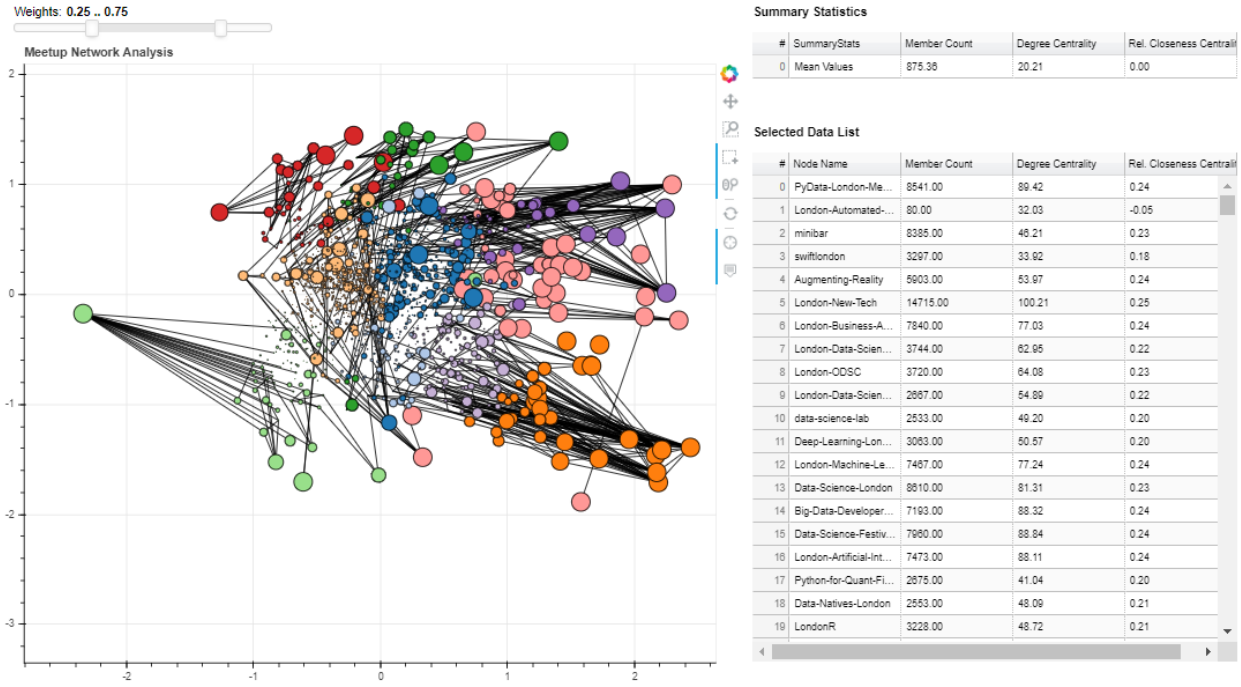
2.2.2 Analytical Task: To build models to understand the key characteristics which drive Meetup group growth and assess the model quality

Percentage based growth metrics tend to be vulnerable to large outliers which are often not meaningful and can damage model performance. Fortunately, visual techniques can be used iteratively to remove outliers such as iterative clustering of line charts and manual removal of outliers. Adding k-means group cluster data from the prior research question (which was determined using visual inspection of clusters) is also expected to improve performance.

3 Analytical Steps

3.1 To describe how group structure varies throughout the network and identify groups with similar network structure

Applying GraphView



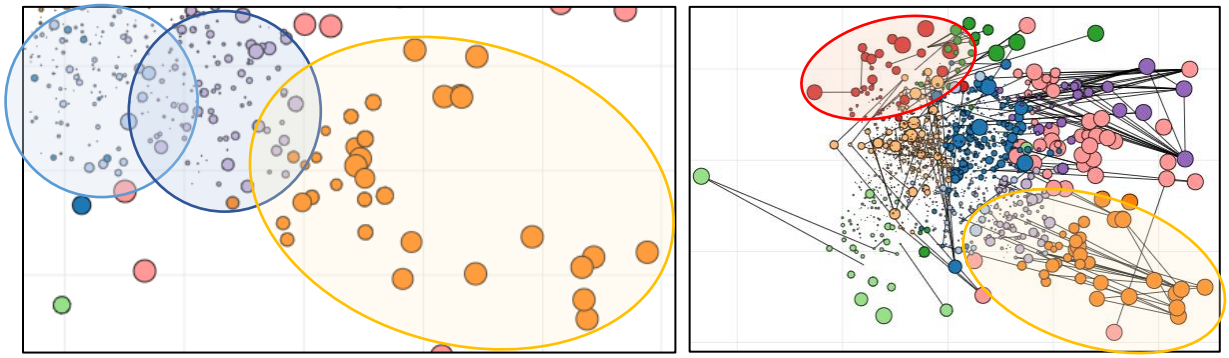
The clusters in the GraphView network can be interpreted through examining the points in a cluster and understanding the commonality between the groups. Reviewing each cluster, it can be determined that the following clusters correspond to different topics.

Summary Data by Cluster

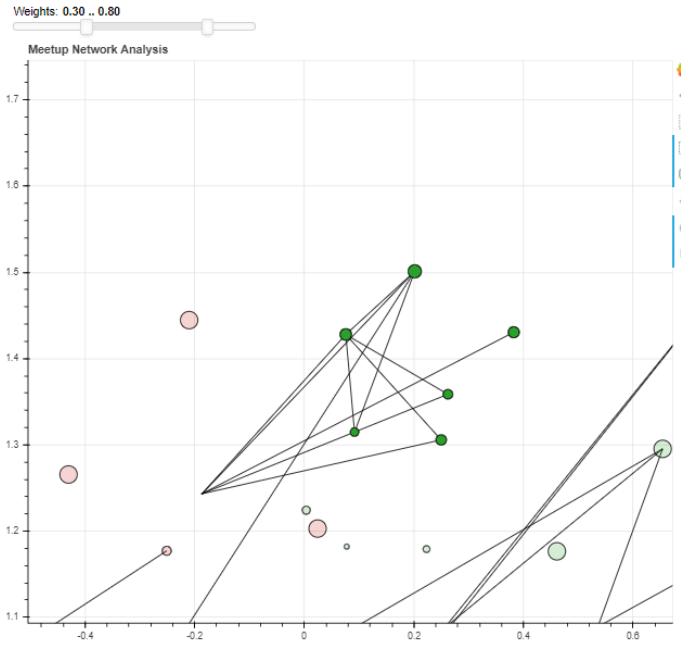
#	Node Color	Assigned Topic	Sample Group	Mean Group Size	Degree Centrality	Rel. Closeness Centrality
0	Dark Blue	CompSci / Coding	Hacks/Hackers London	793	22	5%
1	Light Blue	Niche Data Science	Algorithmic Art	496	19	-1%
2	Orange	Data Science (Academic)	Data Science London	4125	59	22%
3	Peach	Misc. Small Groups	Bioinformatics	400	11	-9%
4	Dark Green	Agile/Scrum	Adventures With Agile	1296	24	5%
5	Light Green	Bitcoin	London Cryptocurrencies	541	21	-1%
6	Red	User Experience	UX Community	1347	20	4%
7	Pink	General Tech	Tech for Good	5457	52	22%
8	Dark Purple	DevOps	London DevOps	1355	31	8%
9	Light Purple	Data Science (Applied)	London TensorFlow	807	30	9%

The academic data science groups have the highest centrality metrics meaning they are important to connecting the network. This can be explained by the relatively large average group size in this cluster and increased interest in data science.

The location is also informative given the approach used. Of note, the data science groups are all very close in proximity and far from user experience / design nodes (red).



The Agile/Scrum network has very strong interconnectedness and given its distance from the data science network, it is unlikely members are part of the data science network.



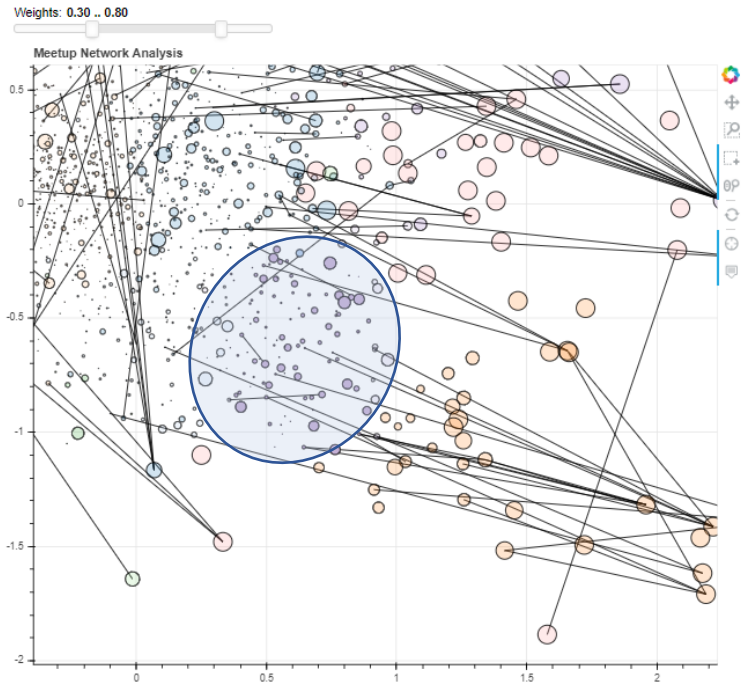
Summary Statistics

#	SummaryStats	Member Count	Degree Centrality	Rel. Closeness Centrali
0	Mean Values	2145.71	30.14	0.14

Selected Data List

#	Node Name	Member Count	Degree Centrality	Rel. Closeness Centrali
0	KanbanCoachingEx...	1960.00	27.81	0.14
1	ACE-Agile-Coachin...	2661.00	28.77	0.16
2	AdventuresWithAgile	3038.00	32.77	0.18
3	London-Scrum-User...	2199.00	32.37	0.16
4	London-Agile-Discu...	2381.00	32.89	0.18
5	London-Scrum-Mee...	2537.00	38.88	0.20
6	LessWorks	244.00	17.87	-0.07

The applied data science group in contrast has proximity to the academic data science network, members in these groups are far more likely to appear in the orange cluster.



Summary Statistics

#	SummaryStats	Member Count	Degree Centrality	Rel. Closeness Centrali
0	Mean Values	777.81	29.60	0.09

Selected Data List

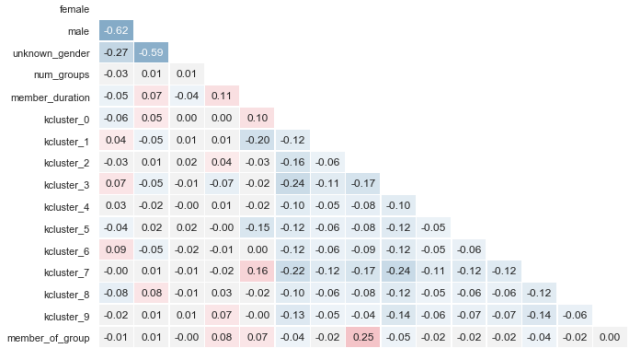
#	Node Name	Member Count	Degree Centrality	Rel. Closeness Centrali
0	London-Automated-...	80.00	32.03	-0.05
1	futureofdata-london	2188.00	41.51	0.19
2	Applied-Deep-Learn...	747.00	31.45	0.12
3	Neurons-AI-UK-Lon...	1818.00	43.93	0.21
4	Enterprise-Artificial-I...	191.00	34.44	0.07
5	Big-Data-Next-Gen-...	1415.00	34.83	0.16
6	datasciencenetworkl...	888.00	33.46	0.13
7	The-Alan-Turing-Ins...	732.00	34.99	0.15
8	opendatascience	305.00	23.38	-0.05
9	London-Get-More-F...	282.00	28.09	0.05
10	Artificial-Intelligence...	692.00	33.68	0.12
11	Data-Science-Busin...	601.00	30.37	0.12
12	DataTech	2114.00	40.33	0.21
13	Data-Engineering-L...	1133.00	39.77	0.18
14	Creative-AI	2322.00	37.50	0.20
15	Applied-Data-Engin...	847.00	30.45	0.14
16	London-Computer-V...	165.00	29.05	-0.01
17	Bayesian-Mixer-Lon...	768.00	25.69	0.10
18	London-Probability-...	115.00	29.12	-0.05
19	AIGLondonDataSci...	729.00	31.71	0.14

3.2 To build models to understand the key characteristics in predicting a given members meetup group membership and assess the model quality

The purpose of the model is to generate a probability that a member will be interested in joining a given group. The target group is “Data-Science-London” and the list can be generated using the probabilities associated with the false positives in the binary classifier of predicting whether a member is part of the target group. The model performance is evaluated based on an f-10 score to ensure it captures enough of the target class while limiting the number of incorrect suggestions.

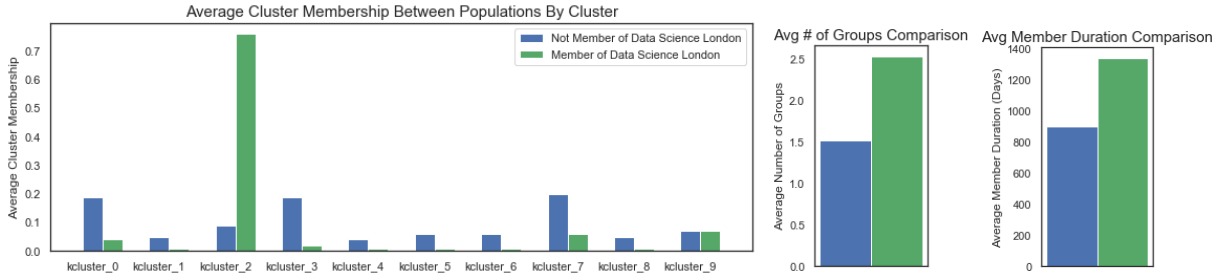
Before selecting algorithms or features, we must first understand the collinearity of the variables using the correlation heatmap.

Correlation Heatmap of Features

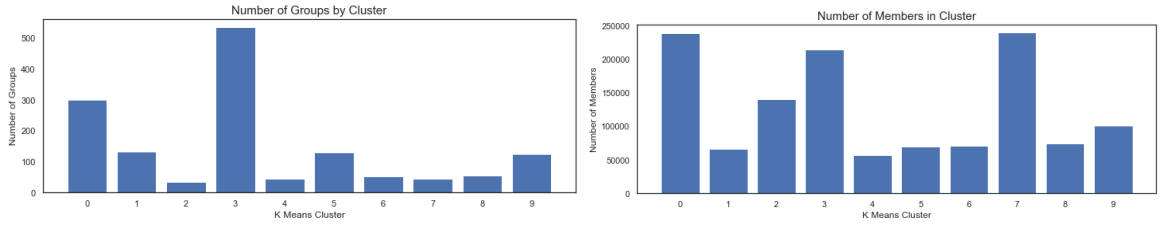


There do not appear to be any significant issues with multicollinearity and therefore all the features can remain in the model. Most variables have a weak correlation with the target indicating we should use a nonlinear model.

The most informative features are likely to be kcluster_2, the average number of groups, and the average member duration based on a comparison of the population means.



The charts below show how the number of groups in each cluster and members in each cluster differ, with the number of members by cluster being more important for predicting group membership.

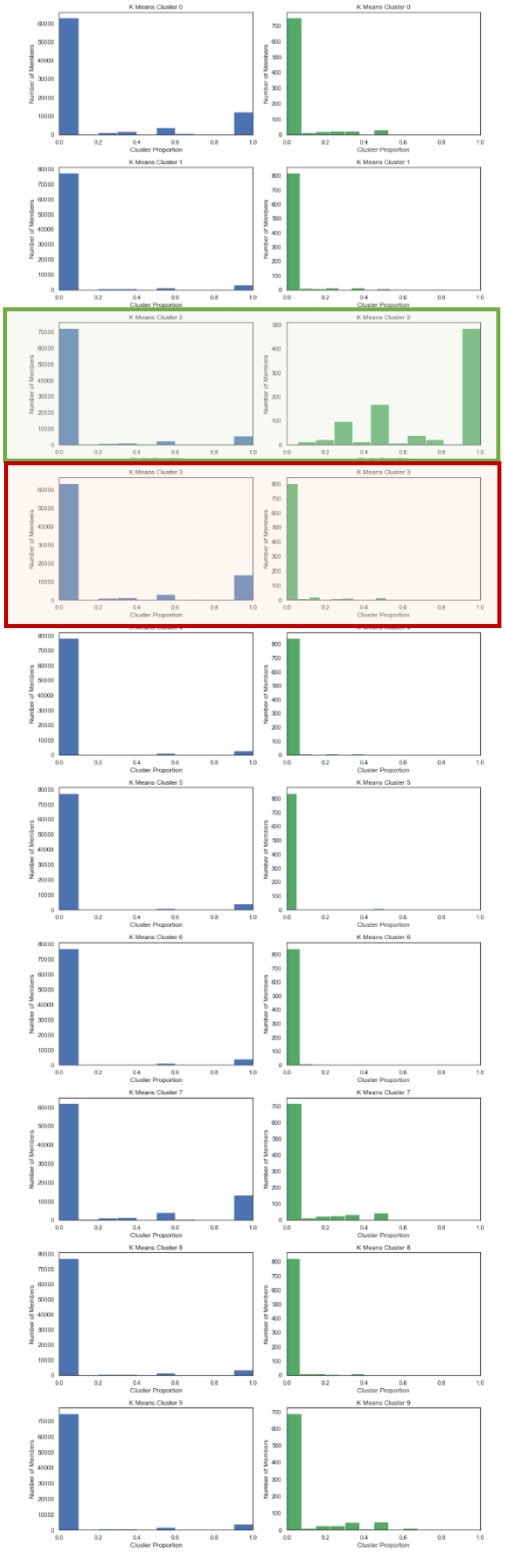


The pair plot below is challenging to interpret even with a sample of data. Instead, if you compare the highlighted distributions to the right, the greater the proportion of membership in cluster 2, the greater the membership in the target group. In contrast, the opposite is shown in cluster 3.

Pair Plot Analysis of Group Cluster Relationships

Groups with a majority of loner members tend to be in cluster 3 as discussed in appendix B.

Distribution of Cluster Membership



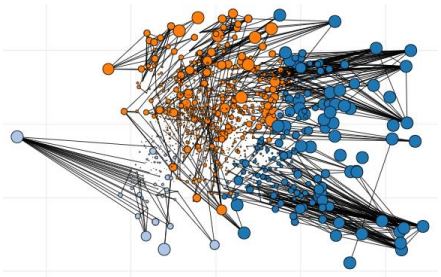
Evaluation of Efficacy of Different Selections for K in K-Means Clustering

Using a combination of a visual inspection of the GraphView network, the model evaluation metrics, and the AUC curves, it is very evident that K=10 is the best hyperparameter setting for k-means.

The f-score metric is the highest for k=10 at 0.52 for LGBM and the AUC is also the highest at 0.97. Testing higher values for k than this resulted in less interpretable / meaningful group clusters.

K-means Cluster (K = 3)

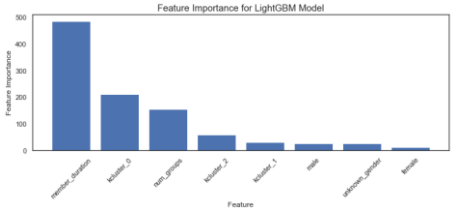
GraphView Network

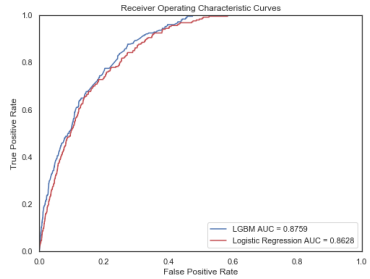


Model Evaluation Metrics

Metric	LightGBM	Logistic Regression
F-score (Beta: 10)	0.3071	0.2915
Precision	0.0494	0.0459
Recall	0.6417	0.6260
AUC	0.8759	0.8628

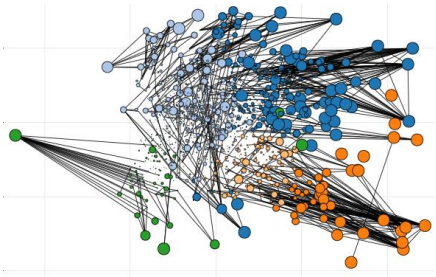
Feature Importance & ROC Curve





K-means Cluster (K = 5)

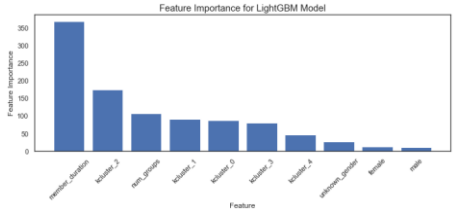
GraphView Network

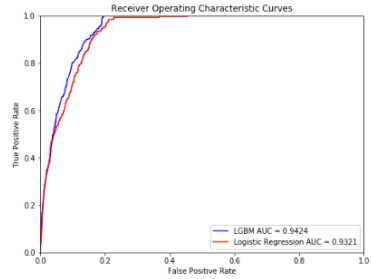


Model Evaluation Metrics

Metric	LightGBM	Logistic Regression
F-score (Beta: 10)	0.4311	0.3850
Precision	0.0769	0.0652
Recall	0.7992	0.7559
AUC	0.9424	0.9321

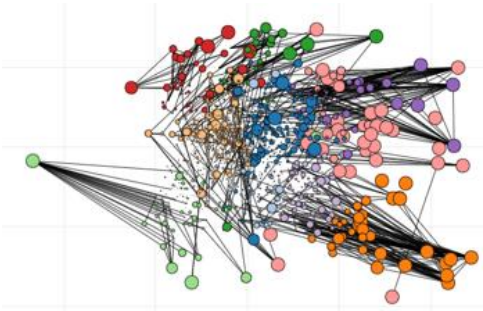
Feature Importance & ROC Curve





K-means Cluster (K = 10)

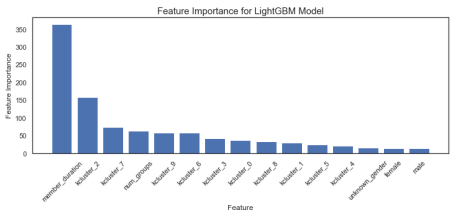
GraphView Network

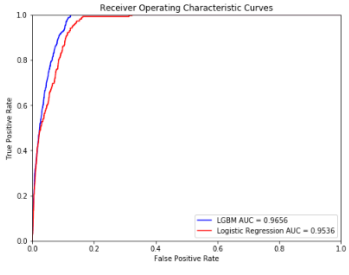


Model Evaluation Metrics

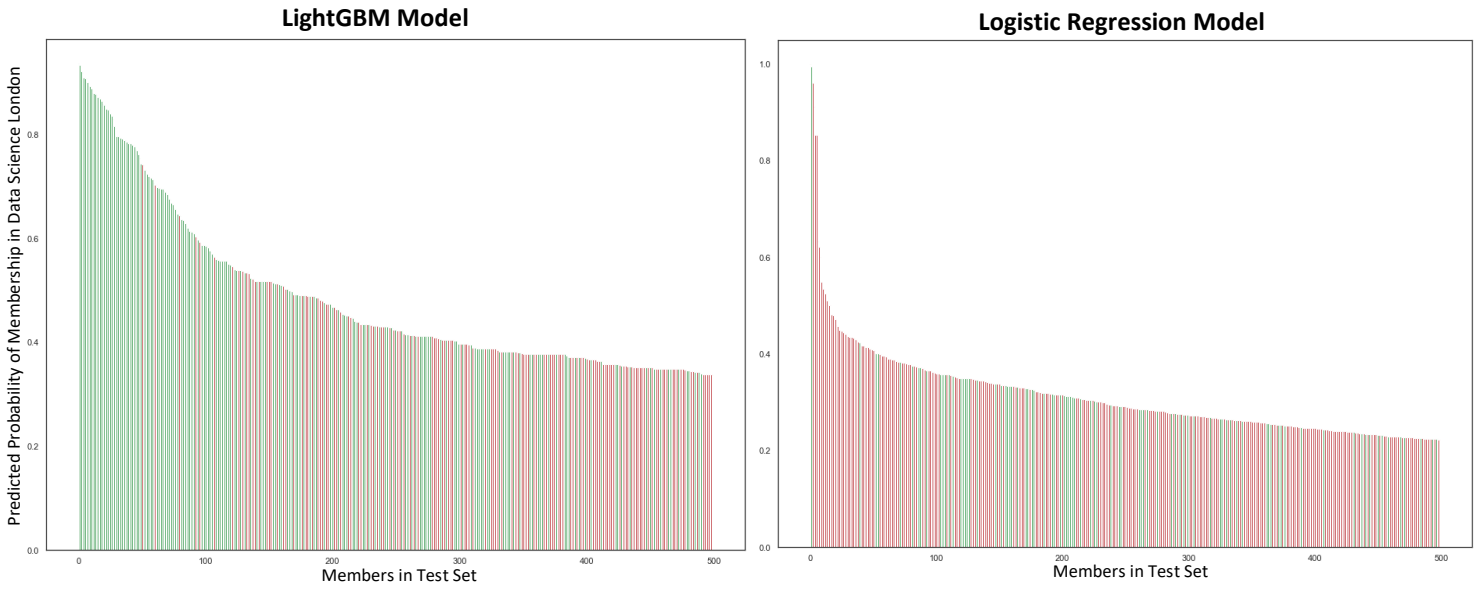
Metric	LightGBM	Logistic Regression
F-score (Beta: 10)	0.5205	0.4596
Precision	0.1177	0.0814
Recall	0.7913	0.8583
AUC	0.9656	0.9536

Feature Importance & ROC Curve

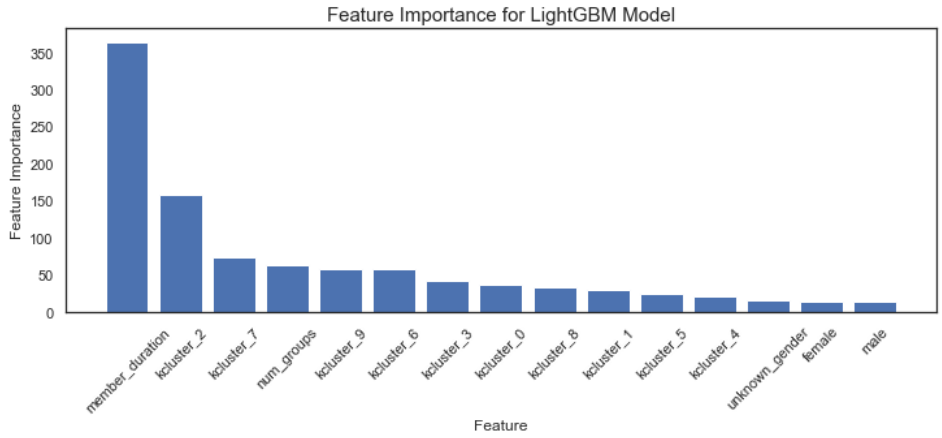




We can further evaluate the performance of the predictions using an Entropy Waterfall. The desired performance would be members with high predicted probability belonging to the group (green towards the left and increasingly red) which is seen in the LightGBM model.

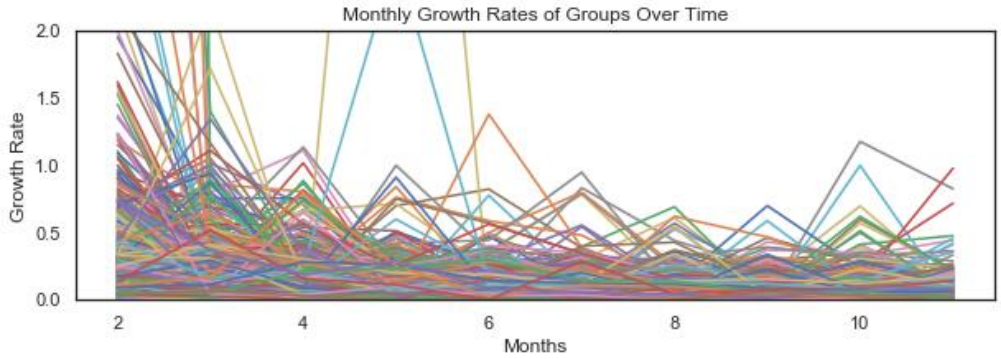


Examining the feature importance confirms that member duration and kcluster 2 membership proportion are very important to prediction.

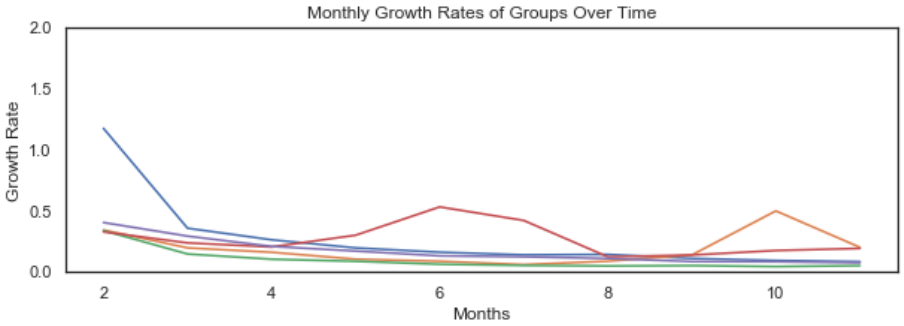


3.3 To describe how group membership changes over time and identify groups with similar growth behaviour

Looking first at the line plots of the growth rate by group over time it is clear there are outliers present which will need to be addressed.



Once outliers are removed, a better approach is to cluster to reveal common growth trajectories. Performance is clustered using k-means and then the means of the clusters are displayed by month.



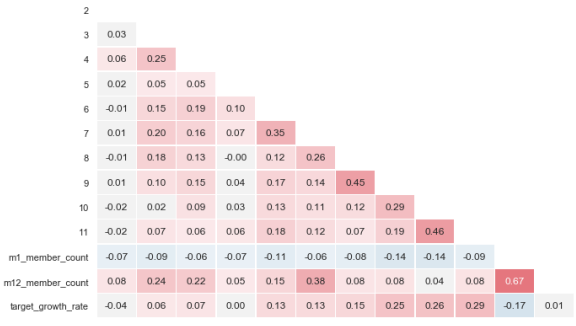
It is evident that most of the clusters follow a similar trajectory with the exception of the red and orange clusters. This may be meaningful in predicting the growth trajectory one-year later.

3.4 To build models to understand the key characteristics which drive Meetup group growth and assess the model quality

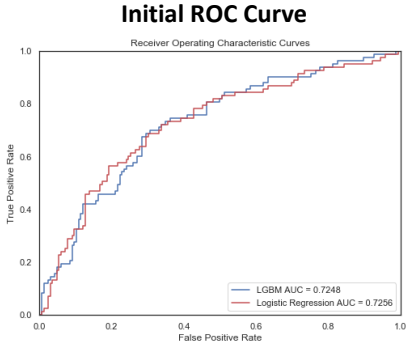
The model is a binary classifier predicting whether a group will double in size in its second year of operation based on the growth trajectory in the first year. The groups are a subset of the technology groups which have existed for more than 2 years.

Before modelling, we need to examine the collinearity for feature selection and inform model selection. No features have problematically high correlation and therefore none should be removed. Certain features have reasonable correlation with the target and given the small dataset we will focus on using logistic regression as it is a simple linear model.

Correlation Heatmap

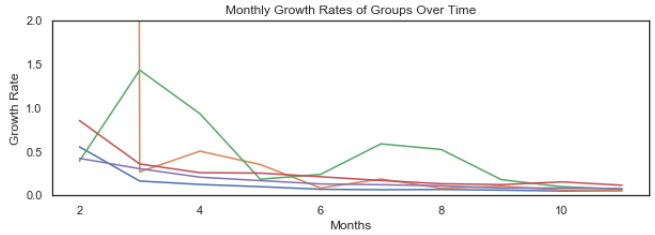


Initial ROC Curve

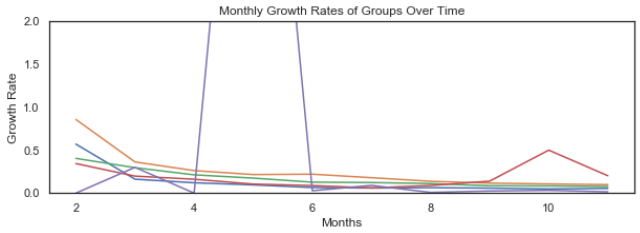


A preliminary model with the standardized inputs achieves accuracy of 69.5% with logistic regression as seen in the ROC curve above.

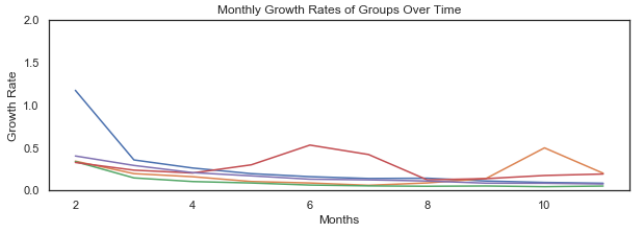
To identify outliers, I clustered the lines using k-means with 5 clusters and removed outliers visually and then re-clustered and repeated until stable.



Removed the training data associated with the orange and green clusters above



Removed the training data associated with the purple cluster above



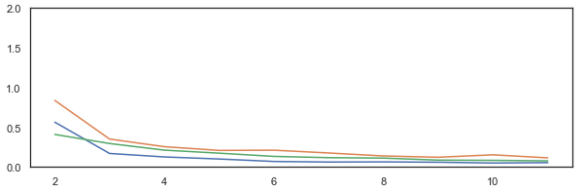
The clusters appear to have stabilized and we can now test performance on the revised training set

The logistic regression test accuracy improved to 71.9% due to removing outliers and now we can test different clustering values for k-means.

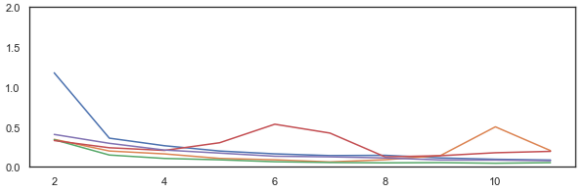
Based on a visual inspection of the charts, a k-means cluster of 3 results in the best performance given there does not seem to be a significant difference between additional clusters.

To further improve the model performance, we can add the group clusters from section 3.2.

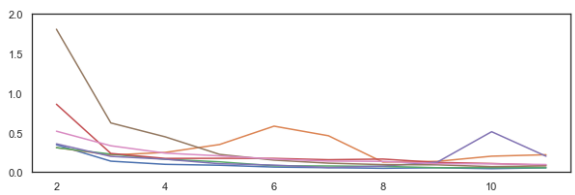
K-means Cluster (K = 3) | Accuracy: 72.7%



K-means Cluster (K = 5) | Accuracy: 71.5%

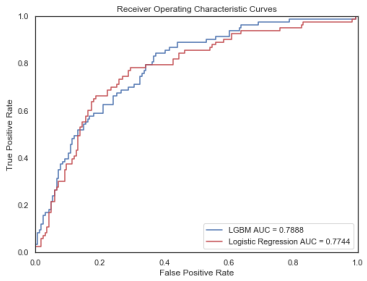


K-means Cluster (K = 7) | Accuracy: 72.3%



This results in a further improvement in accuracy to 74.3% demonstrating there is important information in the group cluster membership.

The increased performance in ROC curves demonstrates the importance of using visual analytics in modelling.



4 Findings

4.1 You can use information from the network graph to determine which members are most likely to be interested in a given group

Using false positives from predictive models was successful in generating a list of members who are potentially interested in Data Science London to whom Meetup can send recommendations. This approach can be generalized to any group in the network and is helpful in understanding the drivers in group membership. Group cluster membership is very important in predicting membership for a given group. Features which are not directly related to the group, such as the number of groups a member is in and the duration of their membership, are also very important specifically to predicting membership in Data Science London although this may not necessarily generalize.

There are indeed differences in the structures between different meetup groups with interests such as Agile/Scrum having very defined shared membership with minimal overlap with other groups. Similarly, it is interesting to note that three different data science related clusters were identified which all had noticeably different membership characteristics.

GraphView is useful for understanding the structure between groups and visually inspecting the efficacy of the clustering algorithms. Interpretability of network graphs deteriorates quickly as the size of the graph increases. Multidimensional scaling was effective for representing group joint membership. The significance of this is elaborated on in the critical reflection section however it is important to note that examining the relationships of complex networks is common problem and creating a repeatable visual analytics strategy is very important. Entropy Waterfall were effective in supplementing raw performance metric scores to determine appropriateness of model predictions.

4.2 You can use information from the network graph to increase accuracy in predicting membership growth

Using the information from the clusters based on the shared membership proportion helped improve the accuracy of predicting whether a group would double in size in its second year which demonstrates the differences in growth trajectories of different meetup group topics. Growth in the later months of the first year is also predictive of whether a group will double in size and groups with higher consistent growth are more likely to double than those with high initial growth but relatively lower growth thereafter during the first year.

Applying visual analytics techniques helped improve performance through both removing outliers and clustering the time series data. Specifically, iteratively applying k-means and identifying outlier clusters

for removal was a very effective strategy which could be repeated in other instances. The relationship between growth in the second year and the growth trajectory in the first year was surprisingly linear with logistic regression showing strong results.

5 Critical Reflection

5.1 Implications of Findings

The findings are very generalizable to different applications within social networks and particularly those with groups as nodes. The recommender system methodology can be applied to Facebook groups, university societies, LinkedIn groups and any situation involving recommending groups to users where one has at least the minimal information of which groups members are currently in. It can be assumed that the prediction accuracy would increase further with increased demographic information, topics the user is interested in, and other data although the power of this approach is that it does not rely on that significant knowledge about the user in making its recommendations. This can be particularly valuable to third party applications without access to significant member data and marketing agencies with incomplete information.

Some more technical findings also have implications to others addressing similar problems to streamline the analytical process. Multidimensional scaling was the most effective approach for representing points versus PCA as it maintains the distance metric. LightGBM performed better than logistic regression which is unsurprising given it can capture non-linear patterns in the data. Other practitioners facing a non-linear relationship should consider LightGBM or other gradient boosted decision trees.

The implications of understanding the characteristics of groups which are likely to double in size in the second year is that meetup groups seeking to grow considerably should focus on maintaining consistent growth. Elevated growth near the end of the first year of operation is indicative of doubling in the following year given the challenges in maintaining membership growth. The implication of group cluster membership improving prediction accuracy means that the growth trajectory also depends on the group membership, so meetup managers should be careful to benchmark against meetups in the same cluster.

5.2 How the Data and Visual Analytics Approaches Facilitated Answering the Research Questions

Structuring the membership data in a projected bipartite graph and analyzing the relationships in GraphView was a crucial component in answering both research questions. The important insights around the variation in group structure and measuring the efficacy of the clusters directly addressed the research questions. I was surprised by how poor the scalability of the standard network graphs visualizations was. I spent a long time trying to make the network graph interpretable without clustering or developing a tool although the high number of edges made analyzing even a modest number of nodes infeasible in terms of both computational resources and more importantly clarity.

5.2.1 Can you use information from the network graph to determine which members are most likely to be interested in joining a given group?

Preliminary exploration of the distributions of the populations including the correlation heatmap, bar charts comparing population means, and bar charts showing the distributions of the clusters were very important in understanding whether features needed to be removed due to multicollinearity, which

features to focus on, and which model will likely perform the best. I was surprised by the efficacy of using k-means and MDS to display the proportion of shared membership between groups. Through visualizing the different values for K and different dimensionality reduction algorithms it was quite easy to pick hyperparameters which are meaningful and interpretable.

Visually inspecting the clusters and evaluating performance using ROC curves and Entropy Waterfalls was very informative in choosing hyperparameters which improve performance. The atypical performance metric meant that analyzing accuracy or AUC as is the case in most binary classification tasks was not relevant so using Entropy Waterfalls added a helpful layer of model evaluation.

5.2.2 Can you use information from the network graph to increase accuracy in predicting membership growth?

Using visual analytics approaches of clustering the data and then visualizing the means of the clusters using line charts was very effective. This approach was used to both identify and remove outliers and add features based on which growth cluster a group was a member of. Both techniques resulted in improved model performance. Adding group clusters derived from the prior the research question also improved performance. Using clustering to identify outliers was surprisingly effective as when removing outliers quantitatively it tends to be challenging to set a meaningful outlier threshold. However, when using visual approaches to identify outliers, it is immediately obvious which points may be damaging performance due to poor generalizability.

5.3 Generalizability of Approaches

The approaches used are broadly generalizable to other situations involving social graph networks. GraphView applies to any weighted undirected graph although could be modified quite easily to be able to handle directed graphs. GraphView would still be useful in analyzing unweighted undirected graphs with the exception of the slider at the top which modifies the visibility of the edges. Networks which could be analyzed using GraphView exist everywhere including transportation networks, banking networks, biology, and social networks.

Clustering using K-means is one of the most widely used unsupervised learning techniques and understanding the efficacy of the clusters using visualization is broadly generalizable. Similarly, reducing the dimensionality of the data in order to visualize high-dimensional data in two-dimensional space using Multidimensional Scaling (MDS) is generalizable to any situation where maintaining between object distances is a priority. This approach would not generalize well to data without a meaningful distance function such as categorical data.

Clustering the components of time series data based on a rate of change to improve visibility, interpretability, and prediction accuracy is also generalizable and has applications in any time series problem including finance, healthcare, and information sciences.

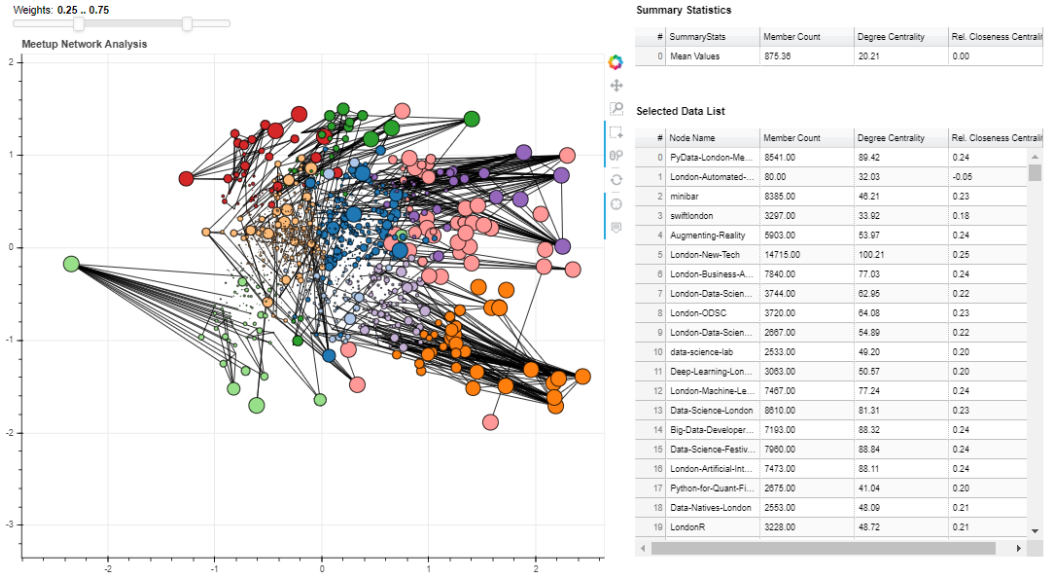
One limitation of the approaches is they are quite resource intensive and would not scale (as currently proposed). Using GraphView as an example, if you were to increase the number of nodes to 100,000 it would be almost impossible to isolate individual groups and their relationship, and it would take significant resources to render. One potential solution is to cluster the nodes in order to improve visibility although information is lost in this process.

Another limitation of the approaches is that it is somewhat sensitive to changes in the membership data therefore these conclusions may not hold through time. The analysis was unable to consider members who left the group which is important for understanding the growth trajectory.

6 Appendix

6.1 Appendix A: GraphView Functionality Overview

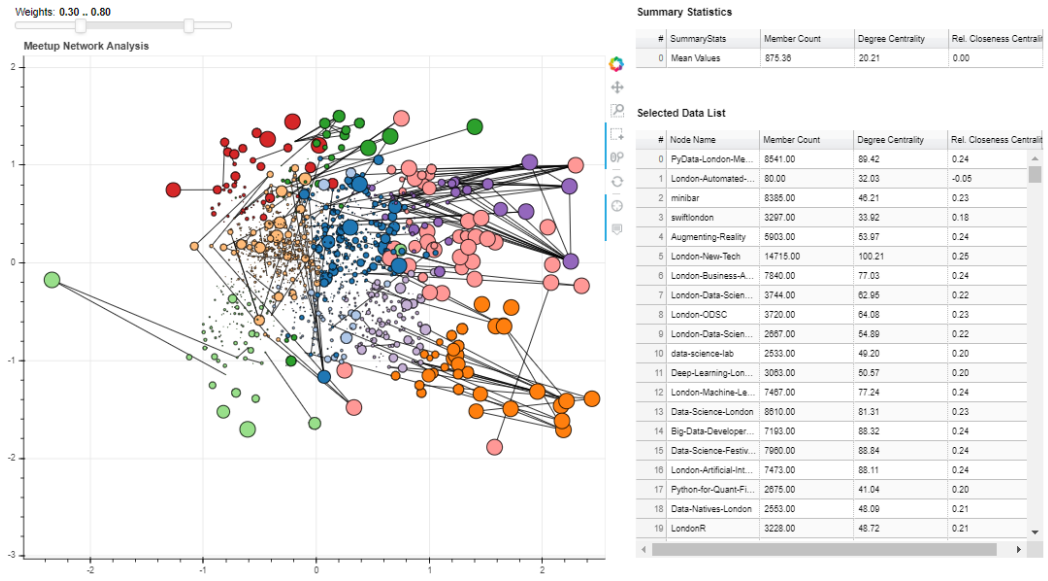
Default Screen



Understanding GraphView:

- Nodes are the different Meetup groups
- Node size is group size (up to maximum threshold)
- Edges are the average proportion of shared membership
- Edges represent two nodes having an average proportion of shared membership within the threshold set using the top slider
- Node position is based on multidimensional scaling based on average proportion of shared membership
- Node color is the k-means cluster derived from shared membership proportion
- Summary Statistics and detailed Selected Data List updates based on selected nodes
    - Degree centrality represents the sum of the weights a node has with other nodes regardless of threshold [22]
    - Closeness centrality measures the mean distance from a node to other nodes [23]. Rather than focus on the absolute closeness centrality, relative centrality (mean centrality of subset / mean closeness centrality of network) was selected as it is easier to interpret

Adjusting the Weight Threshold in GraphView

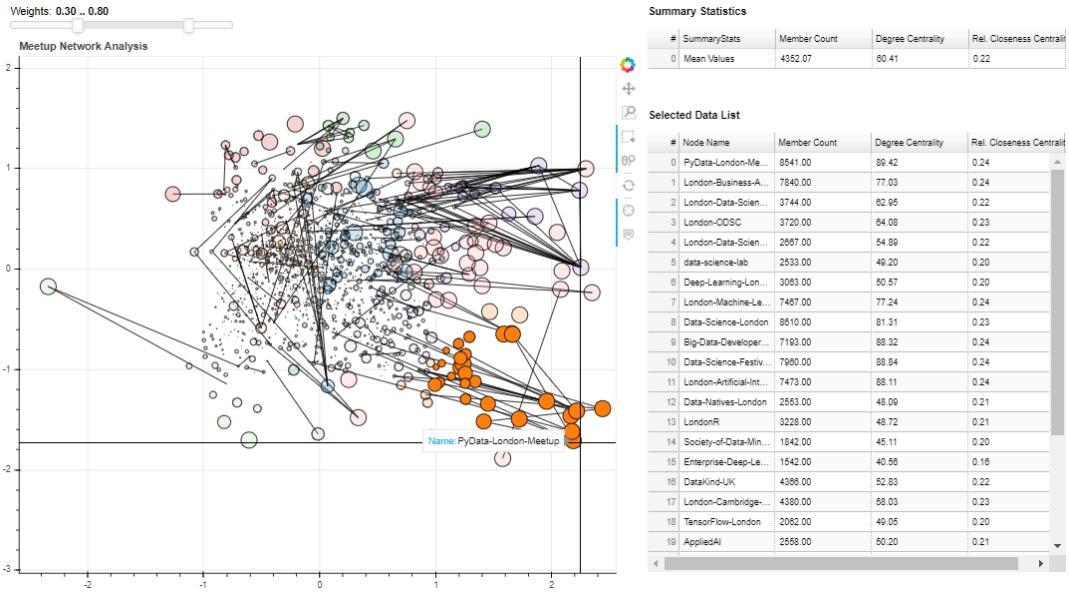


#	SummaryStats	Member Count	Degree Centrality	Rel. Closeness Centrality
0	Mean Values	875.36	20.21	0.00

#	Node Name	Member Count	Degree Centrality	Rel. Closeness Centrality
0	PyData-London-Me...	8541.00	89.42	0.24
1	London-Automated-...	80.00	32.03	-0.05
2	minibar	8386.00	46.21	0.23
3	swiftlondon	3297.00	33.92	0.18
4	Augmenting-Reality	5903.00	53.97	0.24
5	London-New-Tech	14715.00	100.21	0.25
6	London-Business-A...	7840.00	77.03	0.24
7	London-Data-Scien...	3744.00	62.95	0.22
8	London-ODSC	3720.00	64.08	0.23
9	London-Data-Scien...	2667.00	54.89	0.22
10	data-science-lab	2533.00	49.20	0.20
11	Deep-Learning-Lon...	3063.00	50.57	0.20
12	London-Machine-Le...	7467.00	77.24	0.24
13	Data-Science-London	8610.00	81.31	0.23
14	Big-Data-Developer...	7193.00	88.32	0.24
15	Data-Science-Festiv...	7960.00	88.84	0.24
16	London-Artificial-Int...	7473.00	88.11	0.24
17	Python-for-Quant-Fi...	2675.00	41.04	0.20
18	Data-Natives-London	2553.00	48.09	0.21
19	LondonR	3228.00	48.72	0.21

The figure above demonstrates the improved clarity of showing only edges with weights greater than 0.30 by adjusting the threshold at the top versus the prior default figure

Selecting a Subset of Nodes for Examination in GraphView



#	SummaryStats	Member Count	Degree Centrality	Rel. Closeness Centrality
0	Mean Values	4352.07	60.41	0.22

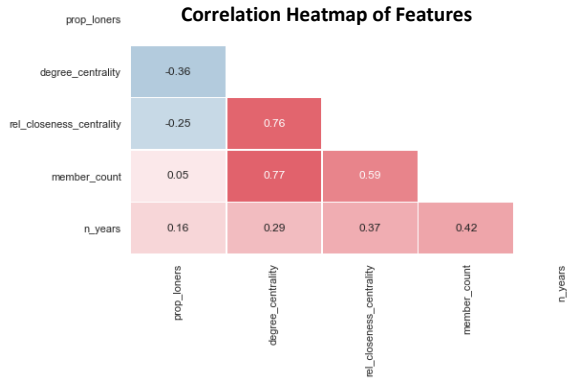
#	Node Name	Member Count	Degree Centrality	Rel. Closeness Centrality
0	PyData-London-Me...	8541.00	89.42	0.24
1	London-Business-A...	7840.00	77.03	0.24
2	London-Data-Scien...	3744.00	62.95	0.22
3	London-ODSC	3720.00	64.08	0.23
4	London-Data-Scien...	2667.00	54.89	0.22
5	data-science-lab	2533.00	49.20	0.20
6	Deep-Learning-Lon...	3063.00	50.57	0.20
7	London-Machine-Le...	7467.00	77.24	0.24
8	Data-Science-London	8610.00	81.31	0.23
9	Big-Data-Developer...	7193.00	88.32	0.24
10	Data-Science-Festiv...	7960.00	88.84	0.24
11	London-Artificial-Int...	7473.00	88.11	0.24
12	Data-Natives-London	2553.00	48.09	0.21
13	LondonR	3228.00	48.72	0.21
14	Society-of-Data-Min...	1842.00	45.11	0.20
15	Enterprise-Deep-Le...	1542.00	40.56	0.16
16	DataKind-UK	4366.00	52.83	0.22
17	London-Cambridge-...	4380.00	58.03	0.23
18	TensorFlow-London	2062.00	49.05	0.20
19	AppliedAI	2568.00	50.20	0.21

The figure above demonstrates GraphView's ability to highlight a subset of nodes to analyze its degree centrality and closeness centrality as well as finding the individual group names by hovering over a node

6.2 Appendix B: Analysis of Loner Members

Certain members are only part of one group in the technology category and it is important to consider applying a separate model only for “loners”. When examining top groups in terms of proportion of loners, it is obvious that the top two are online communities which explains why they would not be involved in other groups.

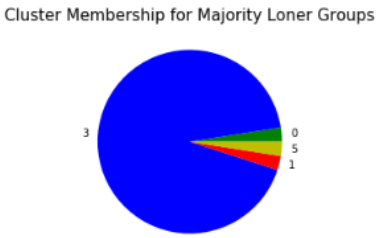
We can then examine the correlation plot between the proportion of members in a given group and other characteristics of the group.



Meetup Groups with Highest Proportion of Loners

	group_name	abs_loners	prop_loners
1	Jenkins-online-meetup	2401	0.87
0	virtualJUG	13451	0.84
355	Association-of-Coaches-Experts-and-speakers	117	0.77
737	LEAGUE-OF-BLOCKS-Quanta-s-world-tour-series	39	0.74
390	London-Workday-HRIS-Meetup	102	0.73
194	SND-London	229	0.71
258	Green-Gathering	170	0.69
114	meetup-group-mjSPBkre	356	0.66
411	London-Decent-bet-Meetup	96	0.66
687	Nintex_UserGroup_UK	47	0.65

The chart to the right shows the cluster membership for groups with over 50% loner members. This demonstrates that the vast majority of loner groups are in cluster 3 which are small miscellaneous groups.



7 References

1. Hempel, J. (2017). WeWork is Buying Meetup Amid an Increasingly Disconnected World. [online] www.wired.com. Available at: https://www.wired.com/story/why-wework-is-buying-meetup/.
2. Weisstein, E. (2018). *Bipartite Graph*. [online] http://mathworld.wolfram.com/. Available at: http://mathworld.wolfram.com/BipartiteGraph.html.
3. Zafarani, R., Abbasi, M.A., Liu, H. & Cambridge Books Online Course Book EBA 2014, Social Media Mining: An Introduction, Cambridge University Press, Cambridge.
4. Zhou, T., Ren, J., Medo, M. & Zhang, Y. 2007, "Bipartite network projection and personal recommendation", *Physical review. E, Statistical, nonlinear, and soft matter physics,* vol. 76, no. 4 Pt 2, pp. 046115.
5. Banerjee, S., Jenamani, M. & Pratihar, D.K. 2017, "Properties of a projected network of a bipartite network", IEEE, pp. 0143.
6. Landesberger, T.v., Kuijper, A., Schreck, T., Kohlhammer, J., Wijk, van, JJ Jarke, Jack, Fekete, J. & Fellner, D. 2011, "Visual analysis of large graphs: state-of-the-art and future research challenges", *Computer Graphics Forum,* vol. 30, no. 6, pp. 1719-1749.

7. Becker, R.A., Eick, S.G. & Wilks, A.R. 1995, "Visualizing network data", *IEEE Transactions on Visualization and Computer Graphics,* vol. 1, no. 1, pp. 16-28.
8. Heer, J. & Boyd, D. 2005, "Vizster: visualizing online social networks", IEEE, pp. 32.
9. Muhongya, K.V. & Maharaj, M.S. 2015, "Visualising and analysing online social networks", IEEE, pp. 1.
10. Gou, L., Zhang, X., Luo, A. & Anderson, P.F. 2012, "SocialNetSense: Supporting sensemaking of social and structural features in networks with interactive visualization", IEEE, pp. 133.
11. Kruskal, Joseph B. "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis." *Psychometrika* 29, no. 1 (1964): 1-27.
12. Buja, A., Swayne, D.F., Littman, M.L., Dean, N., Hofmann, H. & Chen, L. 2008, "Data Visualization with Multidimensional Scaling", *Journal of Computational and Graphical Statistics,* vol. 17, no. 2, pp. 444-472.
13. Haarman, B.C.M., Riemersma-Van der Lek, Rixt F, Nolen, W.A., Mendes, R., xhage, H.A. & Burger, H. 2015, "Feature-expression heat maps - A new visual method to explore complex associations between two variable sets", *Journal of Biomedical Informatics,* vol. 53, pp. 156-161.
14. Bewick, V., Cheek, L. & Ball, J. 2004, "Statistics review 13: receiver operating characteristic curves", *Critical care (London, England),* vol. 8, no. 6, pp. 508-512.
15. DeLong, E.R., DeLong, D.M. and Clarke-Pearson, D.L., 1988. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, pp.837-845.
16. Hanley, J.A. and McNeil, B.J., 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, *143*(1), pp.29-36.
17. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. and Liu, T.Y., 2017. Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems* (pp. 3146-3154).
18. Brownlee, J. (2016).| Feature Importance and Feature Selection With XGBoost in Python. [online] www.machinelearningmastery.com. Available at: https://machinelearningmastery.com/feature-importance-and-feature-selection-with-xgboost-in-python/.
19. F. Tak-chung, “A review on time series data mining” in Engineering Applications of Artificial Intelligence, Volume 24, Issue 1, 2011, pp. 164-181
20. J. J. Van Wijk and E. R. Van Selow, "Cluster and calendar based visualization of time series data," Proceedings 1999 IEEE Symposium on Information Visualization (InfoVis'99), San Francisco, CA, USA, 1999, pp. 4-9.
21. Roelofsen, P “Time Series Clustering”, Master Thesis in Business Analytics, March, 2018
22. Opsahl, T (2017). |Node Centrality in Weighted Network. [online] www.toreopsahl.com. Available at: https://toreopsahl.com/tnet/weighted-networks/node-centrality/
23. Degree Centrality [online] Available at: https://www.sci.unich.it/~francesc/teaching/network/closeness.html