**Team Members:** Logan King, Connor Mignone, Ty Painter

1. An updated overview of your project (highlighting any major changes)

At this stage of our project, we have created two graphs for the available data: one for the customers and another for the articles. Moving forward we need to figure out a solution to connect customers to articles. Currently, we have our graphs structured as follows:

***Customer Graph (GC), nodes (nc)-customer, edges (ec)-purchase***
In the customer graph, if customers (node) bought the same article, they are connected (edge)

***Article Graph (GA), nodes (na)-article, edges (ea)-purchase***
In the article graph, if articles (node) are bought by same customer, they are connected (edge)

We did not have any major changes to our approach up to this point. Over the time period since our proposal, we have been developing a method to address our problem of interest. We hope to be able to recommend a set of articles that customers are most likely to buy for our final project deliverable.

2. An updated weekly timeline for the remainder of the semester (highlighting any major changes)

For the remainder of the semester, our timeline is as follows:

**Current -** Graphs created and edges developed from our datasets of interest.
**Week ending 4/17 -** Decide upon methodology for recommendation system and devise strategy to link customer graph to article graph. Submit midpoint check.
**Week ending 4/24 -** Implement recommendation system on dataset, interpret results and fine-tune to improve results.
**Week ending 5/1 -** Create and give presentation, work on final writeup for submission.
**Finals week ending 5/6 -** Execute any remaining work necessary for submission.

3. List challenges you have had so far (at least 2)

We have encountered a few challenges throughout the duration of our project. Our data contained multiple features for both customers (club_member_status, fashion_news_frequency, age, postal_code) and articles (product_type, product_group, graphical_appearance, color, department, index, section, garment_group, etc.), which could be used to represent node features in our graph. We are deciding whether or not to include node features in our graph, as they may be used to improve our final recommendation system, however this would introduce an extra layer of complexity to our methodology. Currently we are leaning towards not including node features.

Additionally, we are deciding which prediction methodology would be the best to use given our problem of interest, accuracy and complexity tradeoff, and time constraint. We discovered a library called TorchRec, a PyTorch domain library for recommendation systems. According to the website, “This new library provides common sparsity and parallelism primitives, enabling researchers to build state-of-the-art personalization models and deploy them in production.” This tool seems like a useful source to build our recommendation system, however it appears to be complex to learn how to use, set up, and integrate into our workflow. We are investigating other methods to develop our recommendation system, such as some of the methods introduced in class and other available recommendation system methods. Our next step will be to decide upon and implement an appropriate method for our problem of interest which also satisfies the constraints of our project.

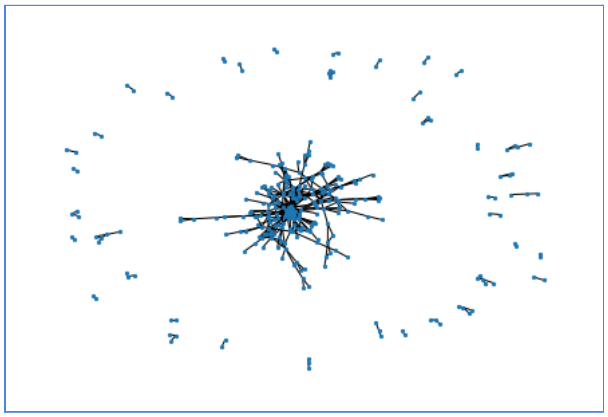
4. Provide basic statistics of the dataset(s) that you'll be using (e.g., number of features, any domain knowledge specifics, number of edges, number of graphs, etc)

There are three datasets. Each can be thought of as a matrix in a matrix factorization problem. The articles of clothing dataset has 105542 rows and 25 columns. The data is fairly self explanatory. All fields could be described by what one might see online shopping, or article characteristics such as ‘t-shirt’, ‘red’. The customer dataset has 1371980 rows and 7 columns. Each row in each respective dataset represents a unique article of clothing or user. The last dataset is the transactions data, which provides customer identification number and article purchased, as well as other possibly useful variables such as transaction date. The data needed minimal, if any cleaning. The largest issue with the data is its size. The transaction data is half a gigabyte when the file is zipped.

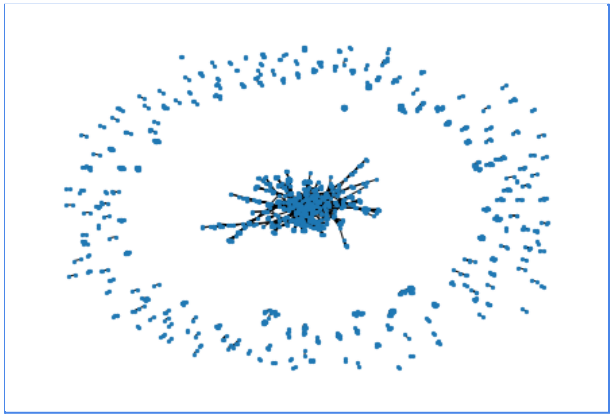
5. Provide at least one table/figure analyzing the data beyond the stats provided above with also mentioning the types of figures/tables that you anticipate including in your project paper/presentation.

As previously mentioned, we created two initial graphs. Both graphs only incorporate the first 2,000 rows of transaction data to reduce compute time and provide a brief summarization. The first graph displays customers as nodes with connecting edges to indicate customers who bought one or more of the same products. The article graph shows articles of clothing as nodes with edges representing if a pair of articles were purchased by the same customer. There are more unique articles of clothing (1,496) compared to unique customers (599).

*Customer*



*Articles*



6. If you have not done so before in your project proposal, please provide a clear discussion of the methodology in terms of the modeling that you'll be using (if this is included in your project, which it seems to be more the large majority)

As discussed in the challenges, we hope to use TorchRec but we may need to adjust if the package seems too hard to work with. TorchRec was released last February and does not have a lot of associated documentation. It might be worth considering splitting a very large dataset into smaller more workable sets, construct a smaller model, then consider scalability. Or, simplify a large scale model that has less moving parts and is less prone to data size issues. Regardless, we hope to use some kind of collaborative item-based or user-based filtering. User might be more  appropriate as the ultimate goal is to predict article IDs for users. Matrix factorization would also be great for this problem. Our greatest concern which we will have to adjust for is scalability. Currently, we built an item-based collaborative filtering method to test capability and run time.

7. How will you evaluate the project (e.g., if it is a regression problem perhaps you might be using RMSE)

The evaluation is defined by mean average precision of 12 article predictions.

Defined by H&M:

Where $U$ is the number of customers, $P(k)$ is the precision at cutoff, $n$ is the number predictions per customer, $m$ is the number of ground truth values per customer, and $rel(k)$ is an indicator function equaling 1 if the item at rank $k$ is a relevant (correct) label, zero otherwise.

$$MAP@12 = \frac{1}{U}\sum_{u=1}^{U}\frac{1}{min(m,12)}\sum_{k=1}^{min(n,12)}P(k)\times rel(k)$$