Modeling The Connection of Products, Demographics, and Social Media

Wyatt Mayor
University of Illinois Urbana-Champaign
Champaign, USA
wpmayor2@illinois.edu

John Lay
University of Illinois Urbana-Champaign
Champaign, USA
johnlay2@illinois.edu

Prathamesh Nadkarni
University of Illinois Urbana-Champaign
Champaign, USA
pyn2@illinois.edu

ABSTRACT

Extracting valuable insights from complex datasets has become a necessity for Social Media Marketing. Given the vast user base on social media platforms, spanning millions in numbers, each possessing purchasing potential, understanding these trends is of paramount importance. In this paper, we have proposed a Graph of Graphs framework for data mining, incorporating multidimensional aspects including product categories, demographics (specifically age), and social network platforms. Our approach is summarized as constructing a hierarchical network representation with these three layers, with edges connecting each layer, that is then mined to extract numerical social media insights based on an input product category. This model enables more accurate and insightful analysis, empowering businesses to make informed decisions in a complex and dynamic marketing landscape.

ACM Reference Format:
Wyatt Mayor, John Lay, and Prathamesh Nadkarni. 2024. Modeling The Connection of Products, Demographics, and Social Media. In . ACM, New York, NY, USA, 11 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

A common problem in data mining is the representation of multiple networks merged into a larger network. These networks are separate, but share characteristics, relationships, correlations. Representing these networks as one allows data mining to be performed on the aptly named “Network of Networks" (NoN), discovering relationships between the networks. Forming this NoN requires determining each individual network’s nodes and edges, as well as the edges between network layers. Creation of such edges should be done via an algorithm rather than hand-labeled to maintain generality, and then further algorithms should be used to extract meaningful information from this NoN.

1.1 Intent

Our specific project requires an effective representation of social media, products and their category, and demographics. Our goal is to develop a method for representing these smaller networks into a larger NoN and then mining it to answer the following question: given a product, on which social networks should we advertise? More precisely, given a Product node, return a weighted set of Social Media nodes representing how much “value" an advertisement for this Product on this Social Media would provide. We take the following steps to answer this:

(1) Constructing a Product network $\mathcal{P}$, a Demographic network $\mathcal{D}$, and a Social Media network $\mathcal{S}$
(2) Determining edges and weights from $\mathcal{P}$ to $\mathcal{D}$ and $\mathcal{D}$ to $\mathcal{S}$; note there need not be backwards edges since our mining is unidirectional.
(3) Traversing this layered Network of Networks to find the weighted set of Social Media nodes that represent advertising value for this product on each social media network A

The networks can be built using proprietary user-provided datasets. We use public domain datasets in our implementation, but note this paper describes and provides a general algorithm, meaning a user can simply swap the datasets for their own for our code to pull from, and a more informative or specialized network model can be created. A generic network model with some sample nodes is shown in figure 2.

Figure 1: Symbols

Symbol	Definition
$\mathcal{N}$	The multilayered network
$\mathcal{P}$	Product network
$\mathcal{D}$	Demographic network
$\mathcal{S}$	Social Media network
A	Advertising values
V_G	The nodes of network G
E_G	The directed edges of network G
$w(v_1, v_2)$	Weight of directed edge (v_1, v_2)

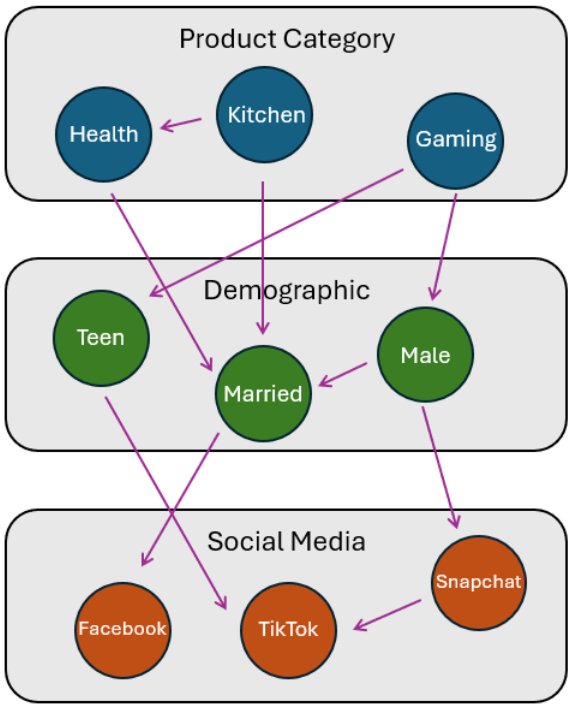
2 NETWORKS

We propose a weighted and directed network with three layers: Product, Demographic, and Social Media. Within each network, there will be directed edges inside corresponding to their derived correlation (we expect a larger edge from Kitchen to Health than from Gaming to Kitchen). Between each network will be similarly weighted edges (we expect a larger edge from Teen to TikTok than Senior to TikTok). Edge weighting is defined in a probabilistic manner, such as the weight from Teen to TikTok can be considered the probability of a user being on TikTok given that they are a Teen,

CS 512, May 2024, Urbana, IL, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/XXXXXXX.XXXXXXX

Figure 2: Network Structure

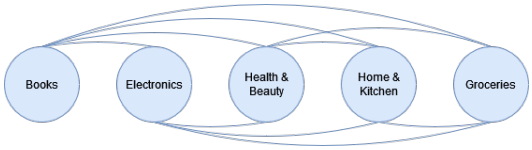


or $w(Teen, TikTok) = P(TikTok|Teen)$. While in theory this implies the network can be fully connected, if desired, some threshold can be observed before creating an edge to prevent uninformative connections and preserving interpretability. Mining will then be performed in a downward fashion with respect to figure 2, starting at some input node p in the Product network, tracing down to our weighted set advertising value A. Note that the nodes displayed in figure 2 are not what we implement, as the datasets we use lack the necessary features. However, the main objective of this work is not the specific network we create in code, but algorithms to produce such a network given a set of appropriate and valuable datasets.

2.1 Product Network

The Product network $\mathcal{P}$ is a small network comprised of a variety of product categories of arbitrary granularity, such as Kitchen, Gaming, and Beauty. In our implementation, we use the five main categories health and beauty, books, groceries, electronics, and home and kitchen for the product nodes which is shown in figure 3. These five categories are a smaller proof of concept to a larger network that could be used. The idea is to classify, either manually or via intelligent systems, a specific product (such as a knife set) into an established product node category (Kitchen). A company with a large amount of product data could break these categories into smaller subcategories to get more refined data insights from our connected networks.

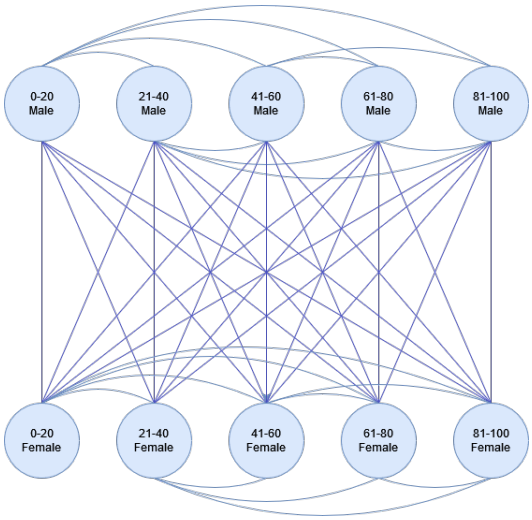
Figure 3: Product Network



2.2 Demographic Network

The Demographic network $\mathcal{D}$ for our example consists of nodes that are age ranges and edges that define the relationship between each age range. The farther the age range, the less related they are. The age ranges are determined by some parameter k. For example, If k is 5 then the age ranges would be 0-5, 6-10, 11-15, 16-20, etc .... Note we are limited to age by our data, but these nodes can be any demographic marker such as location and gender if the user can provide the needed data. Edges between these nodes can more generally be represented as likelihood of belonging to some demographic given another demographic, such as the probability of being a particular gender given that they are college educated.

Figure 4: Demographic Network



2.3 Social Media Network

The Social Media network $\mathcal{S}$ contains a node for each major Social Media network that a company might target, such as Facebook, YouTube, and Snapchat. The edges between Social Media nodes describe how the users overlap between sites. This can be a small set of social media websites that the user is particularly interested in, but is recommended to be broad so that all demographics, even those the user may not initially be interested in, can be properly explained within the network. Our data contains major social media websites used in North America.

Figure 5: Social Media Network

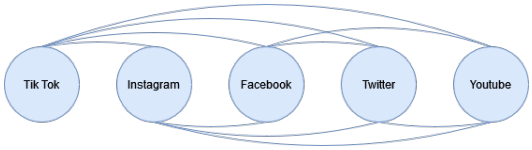
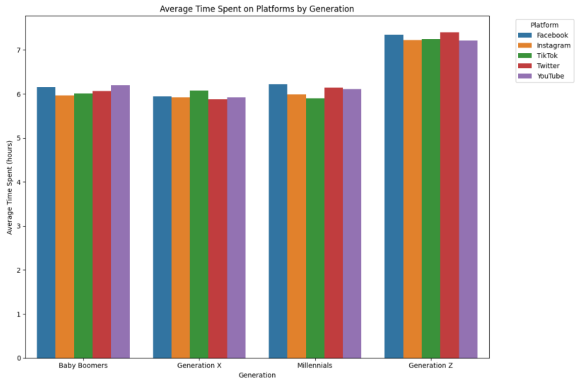


Figure 6: Usage of Social Media wrt Demographics



3 INTERNETWORKING DATASETS

To use the full model, two datasets are needed to establish links between three networks: (1) from Product to Demographic and (2) from Demographic to Social. Here we describe what these datasets might look like as needed for us to extract statistical insights from the relationship between a product category and a particular social media platform.

3.1 Product to Demographic

In order to delineate the connections between nodes within the product network and those in the demographic network, we relied on a comprehensive retail dataset. This dataset needed to encompass distinct transactions, providing details such as the purchased product, its corresponding category, and the age of the customer. This granularity was essential for accurately mapping the interplay between product preferences and demographic characteristics. By incorporating these specific data points, we aimed to elucidate the nuanced relationships between consumer behavior and demographic attributes within our analysis.

3.2 Demographic to Social Media

In order to delineate the connections between nodes within the demographic network and the social media network, we sought out a social media usage dataset to uncover the intricate relationships between demographic segments and their corresponding social media engagement patterns. Specifically, we required data that spanned diverse age groups, ranging from 18 to 70 years, capturing their average time spent on various social media platforms. By obtaining insights into the nuanced digital behaviors of different age cohorts, we aimed to construct a more nuanced and detailed understanding of the interplay between demographics and social media usage.

4 METHODOLOGY: NETWORKS

The three inner subnetworks consist of directed graphs, with each node being a feature, and each edge corresponding to how associated they are, specifically how often they occur together in the data.

4.1 Product Network $\mathcal{P}$

To define $\mathcal{P}$, we consider each product descriptor in the dataset (Health, Book, etc ...) as a node $\in V_{\mathcal{P}}$ and the edges are some correlation. Consider a dataset in the form of a transaction list, where each data point is a list of items/categories. We can then use tools learned from frequent pattern mining to create edges weighted with association rules, as then interestingness measures such as lift and χ^2.

4.2 Product Network Edges $E_{\mathcal{P}}$

One of our challenges that we tackled was how we should go about extracting weights for the product network. Our approach utilized LLMs to produce classifications scores for all of the categories. There have been many industries that have utilized LLMs for data generation, we assume that prompting LLMs to produce reliable classification scores would be a meaningful measurement of correlation between categories based on the large corpus of data that the LLMs are trained on. Due to rate limits and cost of OpenAI's newest GPT 4.5 turbo model, We used the GPT 3.5 turbo. We expect that GPT 4.5 would produce even better classifications scores.

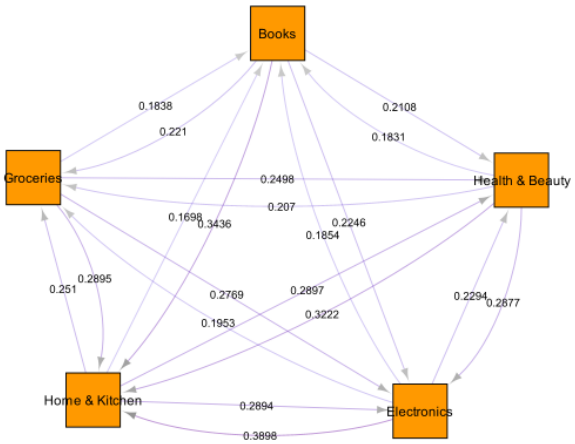
Our group first started by finding a dataset that contained products that closely aligned with the five main categories as previously discussed. We used a large Amazon dataset with over 200k products. We sampled 10k products randomly from this dataset for classification. We assign the product to the category that the product is classified too.

Developing prompts that output parsable snippets can be fairly difficult, but is an essential part of working with an LLM. If the prompting is not very accurate, you run into the problem of wasting a lot of API calls to the LLM, which are limited by OpenAI. To prevent this problem, we defined a question and an output format that the LLM should follow. This method is not 100 percent accurate. We found that around every 10,000 API calls, we get 2,000 that produce unparsable snippets. The 80 percent accuracy works in the confines of our small product dataset but could be costly for a large company with a large product dataset. This accuracy could be improved by using code completion instead of question and format prompting. LLMs have shown very impressive accuracy in code completion. When the LLM responds, we handle any formatting issues by simply reprompting the LLM. This step prevents us from having any corrupt data in the dataset that is created with the classification probabilities.

Finally, to extract the edge weights, we add all the classification probabilities assigned to each category for a specific category and divide by the total number of entries for that category. This gives us the mean of classification probabilities for a designated category. We assign each of these means to their respective edge between two product category nodes. To keep the graphs consistent and provide

better statistical properties, we scale all the edges from a specific category to add up to one by adding all the edges and dividing by 4. The structure of the graph ends up being a bi-directional graph because electronics → books does not have the same relatability score as books → electronics.

Figure 7: Product Network



4.3 Demographic Network $\mathcal{D}$

To define $\mathcal{D}$, we suppose we have a dataset that contains samples in the form of `(self_attribute, [friend1_attribute, ..., friendn_attribute]`. This format can be attained immediately from data such as friends lists on a social website. However, without proper credentials, such data is not easily attainable. Hence, for our purposes, we choose to synthesize the dataset using an attribute of age. We choose age as we can have a strong intuition on what the relations should be between ages, as in we expect a 25-year-old to be more closely aligned to a 28-year-old than a 65-year-old. We use synthesis instead of directly using this intuition to generate edges so that future users with actual demographic data can use our code directly, since the parsing of such a dataset is the same regardless of real or synthetic. The demographic network nodes in general are then the possible attribute values, segmented into reasonable groupings when the potential values are continuous or just unnecessarily granular. Hence in our specific case, we use age intervals of length four years, meaning our nodes are `18-21`, `22-25`, `... 66-69`.

4.3.1 Demographic Network Dataset Synthesis. We detail how we generated this friends list structure. We follow the intuition that people share similar interests with people of similar ages. Hence, to generate a friends list, we simply use a Gaussian distribution centered at their age, then place the generated ages into the appropriate buckets. Specifically, we first uniformly pick from 18-70 what age the sample is, call this a. Then we randomly pick how many friends to generate to add some noise, then generate that many values from the distribution $Normal(\mu = a, \sigma^2 = 1)$. The variance was chosen arbitrarily, after a few sample generations. We then clip the age to the range 18-70, as otherwise people near the limits would have far fewer represented friends, but we don't wish to lose their data, and determine which age ranges the generated ages fall into. For example, a resulting sample with user age of $30 - 33$ we used looks like `["30-33", ["26-29", "30-33", "34-37", "30-33", ..., "26-29"`

4.4 Demographic Network Edges $E_{\mathcal{D}}$

As stated, we use a synthesized dataset for our demographics in the format of a friends list. However, the parsing of such a dataset would be the same regardless of the source: for each sample, add an edge of weight 1 from the $self\_attribute$ to each $friend\_attribute$. If there already exists an edge, we add 1 to the edge. We then normalize all outward edges such that the sum of the weights of the outward edges to other demographic nodes is 1. This step allows us to ensure consistency between layers during the query, as all network layers follow this rule.

4.5 Social Media network $\mathcal{S}$

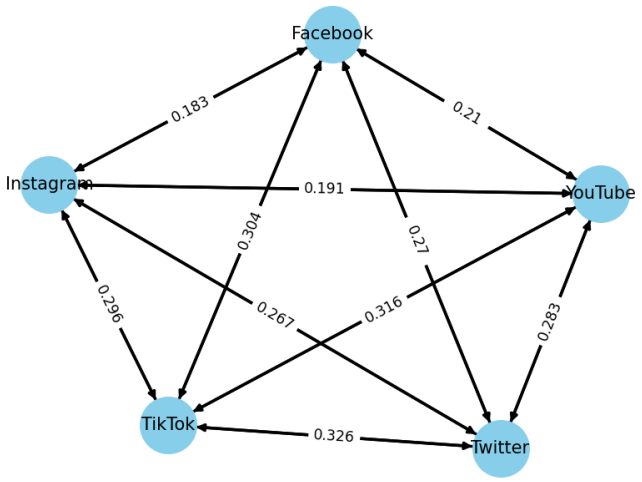
In defining Social Media Network $\mathcal{S}$, our aim was to discern the relative affinity of specific age groups towards different Social Media platforms. To achieve this, we examined the usage patterns of distinct age cohorts across various platforms. By discerning which age groups allocate more time to each platform and how these preferences evolve across different age brackets, we gain valuable insights into the nuanced dynamics of digital engagement. For instance, our analysis reveals that an 18-year-old is more inclined to utilize Instagram, TikTok, and YouTube compared to other platforms, whereas a 35-year-old is more likely to frequent Twitter, Facebook, and YouTube. Similarly, individuals aged 65 and above exhibit a preference for Facebook over other platforms. This detailed exploration allows us to delineate the evolving trends in social media usage across different age demographics, providing valuable context for understanding the intricate interplay between age and platform preference within Social Media Network $\mathcal{S}$. By employing frequent pattern mining algorithms in conjunction with classifiers, we can effectively construct the Social Media Network $\mathcal{S}$.

To develop this framework, we employed a Random Forest Classifier to categorize age groups from age 18-70, into specific subcategories. This approach helps understand the correlations between social media platform usage and distinct age demographics, ultimately contributing to the construction of Social Media Network $\mathcal{S}$. By leveraging this classification technique, we gain valuable insights into how different age groups engage with various social media platforms, thereby creating the structure and dynamics of our network.

4.6 Social Network Edges $E_{\mathcal{S}}$

Utilizing the 'Average Time Spent By A User On Social Media' dataset alongside the established graph structure, we can delineate the edges of the Social Media Network $E_{\mathcal{S}}$. Within this network, each node represents a distinct social media platform, from TikTok, YouTube, Twitter, Facebook, and Instagram. The connection edge between two platforms signifies the probability that a user from

Figure 8: Social Media Network



one platform will also engage with the other. These probabilities undergo normalization, ensuring that the sum of weights for outward edges to other social media nodes equals 1. These normalized values then serve as the basis for determining the weights within the network.

5 METHODOLOGY: INTERNETWORKING

Here we describe how we define two sets of edges between the three networks with respect to our actual code and datasets.

5.1 Connections between the Product Network and the Demographic Network: $E_{\mathcal{P}\to\mathcal{D}}$

We create the edges between the Product network and the Demographic network using our retail dataset. As previously stated, this dataset contains transactions at four different retail locations. These transactions have a lot of different features, including age of the customer and the product that was bought (category). These features allow us to connect the nodes of the two networks. In our extractions process, we divide the number of transactions of a specific age group to a specific product category. Then, we divide by the total number of transactions that have that specific product category. This is essential to keep the statistical properties by scaling all the edge weights between 0 and 1. These edges must be defined as uni directional to keep the statistical properties. For, the purpose of our network this make sense as the idea of connecting these networks is to go from product to the best platform to sell the product on. This could be expanded to go into the other direction by doing the same algorithm in reverse order. However, this is not required for the direction of our project.

5.1.1 $E_{\mathcal{P}\to\mathcal{D}}$ Dataset analysis. As we explored the retail dataset we were using, we realized that the data is pretty evened out. It seems they curated the dataset to have a fairly even distribution of transactions per age. We believe If we were able to collect data like a large retailer, the diversity and depth of the real world dataset would improve the edge weights significantly.

5.2 Connections between the Demographic Network and the Social Network: $E_{\mathcal{D}\to\mathcal{S}}$

For this set of edges, we assume that we have a dataset in the form of "user age : list of social media websites they use". We chose to support this structure as it is a common survey question that can be presented and should be a reasonable dataset for an advertising agency to obtain. However, again we unfortunately do not have such credentials and while we did find multiple data analysis regarding this exact survey question, the raw data was not attainable. Hence we use this data analysis to synthesize data that could have determined these outcomes, meaning despite the synthesis we are still using a real dataset. The details are explained in the following sub-sub-section. For the actual edge creation, we again perform data mining via counting. That is, for each sample, we decide which age range (demographic node) they belong to, then for each social media in their list, we add an edge of weight 1 from their demographic node to the corresponding social media node. If there already is an edge, we add 1 to it. We then normalize the outgoing edges such that they sum to 1.

5.2.1 Demographic to Social Data Synthesis. We use data analysis by Pew Research center as a seed. The data is in the format of "age range: percentage of age range that report using (YouTube | Facebook | Instagram | ...)". To form the data in the format of "user age : list of social media websites they use", we first generate a random age, then for each social media site, we add this website to the list with the probability reported in the seed. For example, for users age 18-29, Pew Research reports 95% percent reported using YouTube, 36% reported using Reddit and so on. Hence if we decide to generate a user with age 22, we add YouTube to their list with probability 0.95, add Reddit with probability 0.36, and so on. This gives a dataset that maximizes the probability of the resulting data, hence we consider this synthesized dataset statistically accurate to the real-world data.

6 QUERY ALGORITHM

Now that we have constructed the full network $\mathcal{N}$, we can perform data analysis on the full graph. As mentioned before, we wish to determine from some product node in the top layer, which social medias on the bottom layer would be most effective to advertise on? To answer this, we perform a weighted traversal on this graph via a modified BFS. That is, from the starting product node, we traverse to both nodes in the same network, such as Beauty will traverse to Electronic as well as a demographic node, each will traverse both laterally and vertically until we reach a bottom layer social node.

The traversal will start with weight 1, then multiply itself with the edges it traverses, meaning longer paths and lower weights will cause this value to diminish more quickly than short paths with high weights. Note in our implementation, we perform these operations under the natural log for numerical stability. We also allow the user to have a hyperparameter λ, which will determine how much we should allow lateral nodes (same network nodes) to influence the final result. That is, if the traversal is lateral, we further multiply the current weight and edge weight by λ. This

is for if a user is not interested in possibly associated users and instead wants to focus on solely the most direct connections.

When a traversal finally reaches a social node, we add its weight to the corresponding value in the output set A, and allow it to traversal to other social medias based on our determined correlations and add to those as well.

Obviously, we will also need some way to prevent infinite cycling, so we use a visited set to prevent cyclic traversals. However, we still need to allow multiple traversals to reach the same node, since otherwise if p is for example Health, and it then traverses laterally to all other products and downward to all demographics, if we have just one visited set, the lateral traversals will never be allowed to reach any demographic node since Health would have added them to visited. Thus we implemented a visited_from set, meaning we only allow one visitation from a parent node to a child node. Thus Health → 22-25 will be permitted, as will Health → Electronics → 22-25, but Health → Books → Electronics → 22-25 will not, since we already have a traversal from Electronics to Health. This by the nature of BFS also means only the shortest parent-child path is permitted, meaning the most direct traversal is the one used.

Algorithm 1: Query

```
Input: A specified product node p and λ ∈ [0, 1] the
       same-network influence
Output: A weighted set of social media nodes A
q = Queue([node=p, weight=1]);
visited_from = Map();
A = Map(v_S: 0 for v_S in social_nodes);
while q do
    v, s = q.pop();
    if v in social_nodes then
        A[v] += s;
    end
    visited = visited_from[v];
    for n in neighbors(v) do
        if n not in visited then
            λ_n = λ if same subgraph else 1;
            visited.add(n);
            queue.append((n, λ_n × s × weight(v,n)));
        end
    end
end
return weights.normalize();
```

7 TECHNOLOGIES

We use Python for the implementation of the model, including data processing and analysis.

We use the NetworkX library for the graph infrastructure. The NetworkX library allows us to easily create and manipulate weighted directed graphs, making the code much more flexible and maintainable than using manual adjacency structures. It also allows us to save our model to a number of text formats, so for models created from particularly large and complex data sets, we can simply save and import the graph for future analysis. It also interfaces directly with our visualization tool, Cytoscape, via a library called py4cytoscape. The visualization tool Cytoscape is an open-source platform for visualizing, analyzing, and modeling networks used to generate our full network shown in Figure 11. It is highly customize and allows dynamic manipulation of the graphs in its application to allow further insight into the model we create. While NetworkX does have its own drawing tool, we found it far too limited for the scale of our graph with many formatting issues that would have been near impossible to adapt to a variety of models, so we felt it worth to include this external dependency.

7.1 Code

The code is presently hosted on GitHub, accessible to the public under the Apache 2.0 License, to preserve copyright and license. This allows individuals to freely modify it to suit their specific requirements and leverage it for their respective purposes. The code is available at:

https://github.com/Prathamesh-Nadkarni/graph_of_graphs.

8 LIMITATIONS

We briefly discuss some potential pitfalls in the model that users should be aware of. The majority of these would arise from issues with the datasets used, as all our model does it attempt to extract the relationships from datasets, so if such datasets are statistically invalid as our model views them, the outputs could be misrepresentation of true relationships.

8.1 Product

The product dataset is limited to the categories that are included in the retail dataset that we use to connect the product network to the demographic network. In a previous section, We discussed that there are five main categories that are included in the retail dataset. In our current approach, we classify the product into one of the five categories. In reality, the depth of products cannot be represented by five categories. This network can be significantly expanded with access to a shop with many transactions. For example, Amazon could collect a dataset by saving the customer's age, the product they bought, and the product category. Amazon has thousands of product categories that are much more specific to products areas. This dataset would allow for a much larger product network with much more specific nodes. This would result in significantly better results for social media advertising values.

8.2 Product to Demographic

In our previous section, we explained that the retail dataset was curated to even out the transactions of age groups for specific categories. Unfortunately, the way we extract our edges relies heavily on an uneven distribution between age groups and products bought. In a real world dataset, there would be a substantial difference between the products bought and the age groups that bought them. This limitation creates edge weights with very slight variation which hinders the strength of the correlation.

8.3 Demographic

Recall that the demographic edges were formed assuming a “friends list" structure, meaning that if a certain demographic was on a lot of friends lists, the edges towards that demographic will be very strong. This means that if you only source from a biased population, the model will be biased towards that population: if you grab the dataset from, for example, Snapchat friends, the vast majority of samples present will be quite young, meaning rarer demographics, say 50+ years of age, will not have a strong link to other people of similar age since by virtue of the platform, the majority of their friends list will be 18-30, leading to a biased model that could be concerning if the product you are marketing is in reality popular among this rarer demographic.

8.4 Demographic to Social

Note that the Demographic to Social edges are meant to represent a normalized $P(S|D)$, that is the probability that a user is active on social media S given they belong to demographic D. Now consider the case where a social media S is popular $\forall D$ and another S' is popular only on a specific D'. If the data shows this, we would have $P(S|D') = P(S'|D')$ and thus the same weights $w(D', S) = w(D', S')$, and we would believe both social medias have equal advertising value for this demographic, thus even for a product that exclusively targets D', their values in A would be similar. However, advertising on the popular S would expose the product to all D as well, who would not be interested, and thus can be considered “wasted eyes". Of course, since all of D' would also see the advertisement, it initially seems arbitrary to advertise on the popular S or the niche S', but the cost of advertising on S' may be much lower simply because it’s more niche, and the edges can’t capture this. We further generalize this in the Query limitation section.

8.5 Social

The social edges represent a spectrum of Social Network platforms, including Instagram, TikTok, Twitter, Facebook, and YouTube. These edges are established based on an understanding of which age groups predominantly utilize each platform. However, age is just one of several factors influencing social media usage. Additional variables such as location, gender, and community dynamics also play significant roles. Presently, our model solely focuses on age demographics, overlooking these other influential factors. To enhance the precision of our model, we can integrate these additional parameters.

In our Social network, we’ve included YouTube, a platform widely utilized across all age demographics, thus attributing it with greater significance compared to other platforms. However, when evaluating platforms of this nature, it’s imperative to normalize the data to ensure fair comparisons. Additionally, understanding the marketing costs associated with each platform is crucial for obtaining valuable insights into their effectiveness and potential return on investment.

Furthermore, expanding the scope to encompass numerous other social media platforms like WhatsApp, Snapchat, Reddit, Quora, and Threads can provide a more comprehensive understanding of the broader landscape of social media usage. This broader perspective is invaluable for refining marketing strategies and gaining deeper insights into consumer behavior across diverse online platforms.

8.6 Query

Consider the case when an input product is popular only among a specific demographic and that demographic does not often use Social Media. While an accurate conclusion might be “don’t advertise on social media for this product”, since our weighted vector is normalized, it may be hard to identify this conclusion since the scale of the ranking values are the same as if the interested demographic is very active on a variety of social medias. Thus the results should be interpreted as “assuming we are advertising on social media, these are the best to do so on", not as “advertising on this social media website is a great idea".

Further, to generalize the discussion in the Demographic to Social limitation discussion, note this model and algorithm does not account for the actual value of advertising on each website. For example, in our data it was unsurprisingly noted that the vast majority of the population uses YouTube, hence it was ranked highly for every product, since if every demographic uses YouTube, if there exists a demographic that is interested in the input product then there exists an interested demographic uses YouTube and thus YouTube scored highly for this product. This does not mean every dollar should be poured into YouTube, since effectively advertising to the full YouTube population might cost significantly more per exposure than other, less popular Social Media sites. Thus the actual cost to advertise on a website can be used to scale the output weights into some A_v, weighted advertising *value* for each website. This is potential future work that can be explored and incorporated in the model given a dataset of advertising costs.

9 EVALUATION

The results presented in Figure 9 demonstrates the query output for all five product categories: Health & Beauty, Electronics, Books, Groceries, and Home & Kitchen. The query traverses through every layer of the network and yields insights into which social media platform holds greater potential for marketing the specified product.

To elaborate further, consider the example of the Books category, which are predominantly consumed by young individuals. In this scenario, the network assigns a relevance score of 0.10 to TikTok and 0.27 to Facebook. Conversely, for Home & Kitchen category, which appeal more to older demographics, TikTok receives a relevance score of 0.08, while Facebook scores 0.31. We show a graphical comparison between the two in Figure 10.

However, it’s important to note that due to limitations inherent in the dataset, YouTube, being a platform utilized by individuals across various age groups, carries greater weight in the evaluation. Consequently, YouTube consistently maintains a higher rank compared to other platforms, owing to its broader demographic reach and usage prevalence.

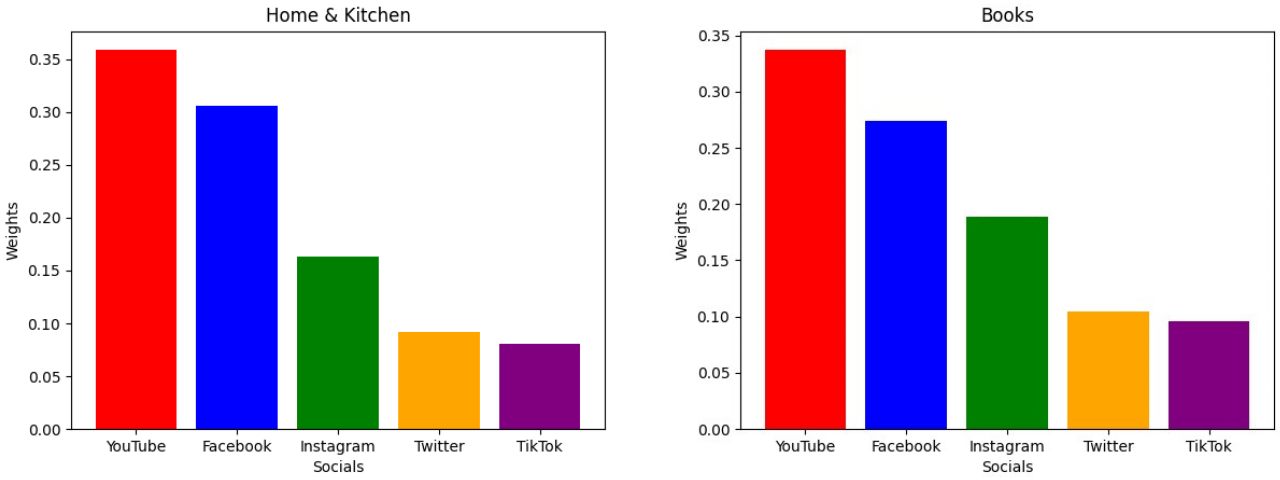
10 CONCLUSION

In conclusion, this research has shed light on the intricate dynamics of product consumption among demographics and their social media usage by platforms. By leveraging sophisticated data analysis

Figure 9: Social Weight per Product Category

Category	Platform				
	YouTube	Facebook	Instagram	Twitter	TikTok
Health & Beauty	0.3485	0.2852	0.1759	0.1014	0.0890
Electronics	0.3465	0.2973	0.1723	0.0938	0.0899
Books	0.3369	0.2744	0.1886	0.1045	0.0956
Groceries	0.3457	0.2864	0.1798	0.1028	0.0853
Home & Kitchen	0.3586	0.3055	0.1635	0.0915	0.0809

Figure 10: Home & Kitchen vs Books output



techniques, including frequent pattern mining algorithms and classifiers, we constructed a comprehensive network using a Graph of Graphs. By embracing a multidimensional approach, we can better navigate the complexities of digital engagement and harness the full potential of social media for targeted marketing and strategic decision-making.

Nevertheless, it's crucial to recognize the limitations inherent in our model. While it offers valuable insights, it falls short in fully incorporating additional influential variables such as location, gender, and community dynamics within the Social Network, and it does not encompass all product categories and subcategories within the Product Network. We have aimed to develop a foundational framework that aids companies in comprehending the intricacies of the social media marketing landscape, provides meaningful insights, and helps understand a more holistic understanding of online consumer behavior.

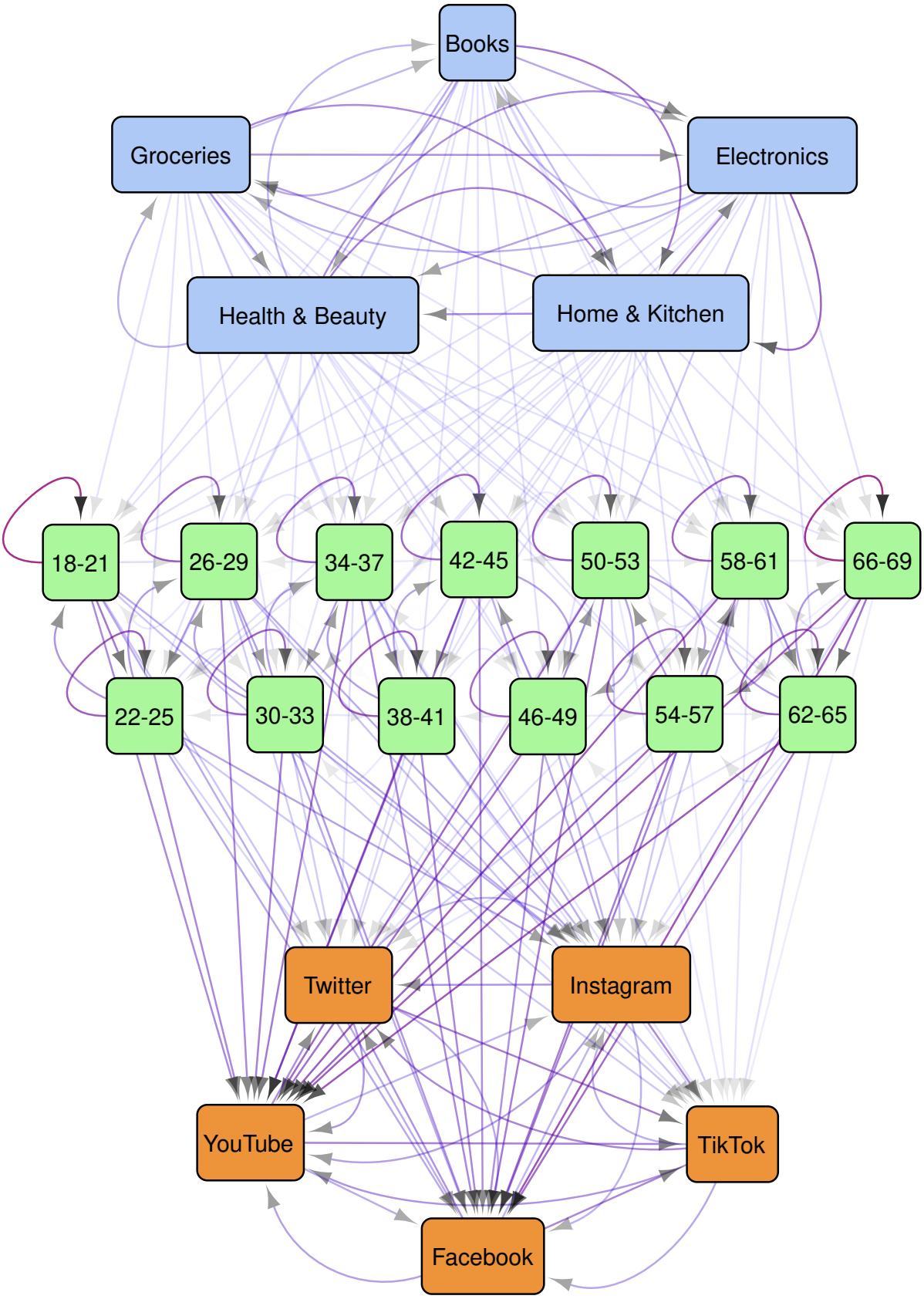
11 FUTURE WORKS

This project has many practical uses. We demonstrated on a small scale how to connect three common networks (product, demographic, and social media). We were limited to our dataset availability. It would be interesting to see the network of network structure that was proposed to be implemented in a large scale setting, by a company with a lot of available data. Another, Future work could explore the impact of expanding the networks, such as adding more components to demographic like age or increasing the product categories. They could compare the advertising values that are mined for specific products to test the effectiveness. Also, We utilized an LLM for edge generation between the product network. Although this is a creative idea, there may be more precise and practical ways to generate the product network edges. There could also be future work in testing the impact of implementing a newer LLM that is more accurate. Finally, expanding this network of networks to have the ability to be mined both ways. Our current setup is designed to be unidirectional between the graphs starting from the product → demographic → social media. There could be more interesting data that is mined from the other direction.

REFERENCES

- Jingchao Ni, Hanghang Tong, Wei Fan, Xiang Zhang: Inside the atoms: ranking on a network of networks. KDD 2014: 1356-1365
- Chen Chen, Jingrui He, Nadya Bliss, Hanghang Tong: Towards Optimal Connectivity on Multi-Layered Networks. IEEE Trans. Knowl. Data Eng. 29(10): 2332-2346 (2017)
- Jundong Li, Chen Chen, Hanghang Tong, Huan Liu: Multi-layered network embedding. SDM 2018
- Yuchen Yan, Qinghai Zhou, Jinning Li, Tarek Abdelzaher, Hanghang Tong: Dissecting Cross-Layer Dependency Inference on Multi-Layered Inter-Dependent Networks. CIKM 2022.
- Jesse N. Moore, Mary Anne Raymond & Christopher D. Hopkins (2015) Social Selling: A Comparison of Social Media Usage Across Process Stage, Markets, and Sales Job Functions, Journal of Marketing Theory and Practice, 23:1, 1-20, DOI: 10.1080/10696679.2015.980163
- Pew Research Center, April 2021, "Social Media Use in 2021"
- Moore, J. N., Raymond, M. A., & Hopkins, C. D. (2015) Social Selling: A Comparison of Social Media Usage Across Process Stage, Markets, and Sales Job Functions. Journal of Marketing Theory and Practice, 23(1), 1–20. https://doi.org/10.1080/10696679.2015.980163
- Fan, Weiguo & Gordon, Michael. (2014). The Power of Social Media Analytics. Communications of the ACM. 57. 74-81. 10.1145/2602574.
- Jamil K, Dunnan L, Gul RF, Shehzad MU, Gillani SHM and Awan FH (2022) Role of Social Media Marketing Activities in Influencing Customer Intentions: A Perspective of a New Emerging Era. Front. Psychol. 12:808525. doi: 10.3389/fpsyg.2021.808525
- J. S. IMMACULATE, A. S. JANET and K. J. C. ANGEL, "A Study of Social Media Analytics," 2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Noida, India, 2021, pp. 1-13, doi: 10.1109/ICRITO51393.2021.9596247.
- Drivas, I.C.; Kouis, D.; Kyriaki-Manessi, D.; Giannakopoulou, F. Social Media Analytics and Metrics for Improving Users Engagement. Knowledge 2022, 2, 225-242. https://doi.org/10.3390/knowledge2020014

Figure 11: $\mathcal{N}$. Opacity and color indicates edge weight.



CONTRIBUTIONS

Wyatt Mayor - wpmayor2

Code:

- Product Network
- Product Network Dataset Selections
- Product Network Edge Custom Dataset Generation Algorithm
- Product to Demographic Edge Extraction Algorithm

Paper:

- Methodology of Product network
- Methodology of Product to Demographic edges
- Limitation of Product Network
- Limitation of product to demographic
- Future Work

John Lay - johnlay2

Code:

- Structure
- Demographic Network
- Demographic Dataset Synthesis
- Demographic to Social Media edges
- Demographic to Social dataset selection
- Query algorithm
- Visualization

Paper:

- Introduction and intent
- Network structure
- Methodology of Demographic network
- Methodology of Demographic to Social Media edges
- Technologies
- Query Algorithm
- Limitations of Demographic
- Limitations of Demographic to Social
- Limitations of Query Algorithm

Prathamesh Nadkarni - pyn2

Code:

- Social Dataset Selection
- Social Dataset Synthesis
- Social Network

Paper:

- Abstract
- Internetworking Datasets
- Methodology of Social network
- Limitations of Social
- Evaluation
- Conclusion