Distribution Agreement

Urshila Choubal April 12, 2022

Ranking Instagram Preferences:
Get to know your friends better through experimental mathematics

by

Urshila Choubal

Manuela Manetta
Advisor

Department of Mathematics

Manuela Manetta

Advisor

Wei Huang

Committee Member

Lori Teague

Committee Member

2022

Ranking Instagram Preferences:
Get to know your friends better through experimental mathematics

By

Urshila Choubal

Manuela Manetta
Advisor

An abstract of
a thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Science with Honors

Department of Mathematics

2022

Abstract

Ranking Instagram Preferences:
Get to know your friends better through experimental mathematics

By Urshila Choubal

Ranking methods offer remarkable potential in creating and revamping recommendation systems. The task of suggesting more relevant and attractive content to users is directly benefited by improving ranking techniques. With graph ranking as the mathematical foundation on which recommendation systems are built, vertex prestige is a critical problem to be addressed. Several models exist that rank vertices in a graph. However, we explore the following methods: HITS, Dominant Eigenvector, and PageRank. We aim to emulate a recommendation system by first gathering primary data from Instagram by tracking the activity of nine participants on the app. With the help of the three ranking methods, we intend to provide our recommendation to the participants based on having accessed their past preferences.

Ranking Instagram Preferences:
Get to know your friends better through experimental mathematics

By

Urshila Choubal

Manuela Manetta
Advisor

A thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Science with Honors

Department of Mathematics

2022

Acknowledgments

I would like to thank several people for their help and support during the production of this thesis.

I would like to thank my thesis advisor, Dr. Manuela Manetta, for her invaluable encouragement, guidance, and optimism. I am so grateful to have had the opportunity to work with her. I would also like to thank Dr. Wei Huang and Lori Teague for being on my thesis committee and encouraging me to pursue mathematics from an interdisciplinary perspective. Finally, thank you to my family, friends, and Peaches without whom none of this would be possible.

Contents

List of Figures

List of Tables

Overview and Motivation

With the overwhelming amount of information available on the internet, there is an urgent need to prioritize and efficiently streamline data for users. Recommendation systems readily solve this problem by analyzing and filtering large volumes of dynamically generated data. Such systems attempt to predict a customer's choices by analyzing their consumption based on past preferences. Since graphs can represent consumption data on the web, the essence of a recommendation system lies in graph ranking and vertex prestige: both of which aim to understand the influence of each vertex in a graph. Thus, the problem now comes down to figuring out which vertex in the graph is of most importance to the user. Vertex ranking methods can numerically make sense of a user's preference, and consequently recommendation systems exploit this technique to provide users with more relevant content. In recent years several ranking algorithms have emerged. However, this thesis focuses on the methods of HITS, Dominant Eigenvector and the famous algorithm PageRank used by Google.

The present work is organized as follows. In Chapter 1, we introduce the reader to the notion of user preferences on social media. In Chapter 2, we build the mathematical foundations necessary to describe the methods used in our experiment. In Chapter 3, we provide the reader with the details of the experiment we conducted in which we design an Instagram account. We then exposed nine Emory students to the account and collected a very specific data set. With the ranking methods discussed in this thesis, we aimed to analyze the Instagram preferences of the participants and provide them with our results. We conclude with a discussion on the results we obtained at the end of the experiment, and we briefly explore possible future directions of research on this topic.

Chapter 1

Social Media and User Preferences

1.1 Recommendation Systems

If you have ever found yourself hopping from one link to another while online shopping, or if you have ever caught yourself scrolling through social media for hours for no reason, it is very possible that the website or app you are viewing may be showing you just the right content to keep you engaged. Gone are the days when online shoppers browsed generic stock for a random item. Today, almost all service providers strive to tailor content for their customers as much as possible to offer them a personalized experience. So how do companies like Instagram, Amazon, Facebook, Tinder, Netflix, and YouTube entice users with relevant content? Of course, data is half of the answer, but you also need some sort of system to narrow down the enormous possibilities that come with gathering massive data sets. That is precisely where recommendation systems come into play. Very generally, a recommendation system filters the collected data and provides tailor-made suggestions to users. This ability to gather users' preferences benefits companies by greatly increasing revenue, customer satisfaction, and customer retention rates. Once the recommendation system is deemed effective - according to specific metrics - it is only a matter of time before the app "learns what you like" and gives you exactly who you need: sometimes even more.

Before diving deeper into the topic that will be explored in this thesis, it is important to broadly understand the two major paradigms of recommendations systems. A general explanation of the two methods will also help ease into the mathematics used to analyse the data collected during the experiment for this thesis.

(**a**) Collaborative filtering methods, for recommendation systems, are solely based on past interactions between users and items in order to produce recommendations. Such interactions are stored in a "user-item interactions matrix". Spotify and Netflix are some of the applications that use collaborative filtering methods. Collaborative filtering can further be divided into

1. Item-Item Collaborative Filtering

2. User-User Collaborative Filtering

(**b**) Content-based methods, unlike collaborative filtering methods, rely on additional attributes of the user and/or item to provide recommendations. User attributes include age, gender, location, job and many more. Let us assume a content-based movie recommendation system as an example.
The attributes of the movie *The Amazing Spider-Man* can include

1. Director - Marc Webb

2. Genre - Superhero

3. Cast - Andrew Garfield, Emma Stone

The recommendation system in question will possibly recommend to any user who watches *The Amazing Spider-Man* movies of a similar superhero genre, movies that feature Andrew Garfield/Emma Stone, or movies that have been directed by Marc

Webb. The more choices the user makes on the platform, by narrowing the user/item attributes, the more targeted results the system will produce.

1.2 Clustering and Ranking

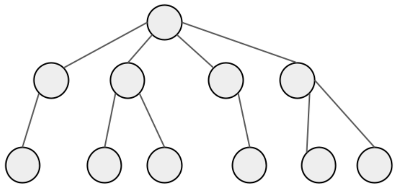
A recommendation system ranks the user's preferences based on a large input data set containing the user's past consumption choices. Exploiting a variety of mathematical models, the system is able to provide a suggestion that it has calculated to be the most relevant to the user. Various models use different methods to rank data in a network. In the following work, we will refer to a network or graph indistinctly. In our case, clustering algorithms can be employed to improve the performance of recommendation methods in social media given the vast quantity of user-generated content that is available online [5]. There are many ways to analyze a graph, which depend largely on the context of the problem we are looking at. Often we try to identify the most "important" data points in our data set: equivalently, the most important vertices in our graph. Note that what we define as an "important" vertex can change with context. Consider, for example, a network generated by the dozens of flights routes of a particular airline. We may want to figure out which city is most visited or least visited in this network. In both cases, the city (vertex) we care to identify will be deemed as "important". With social media, and specifically recommendation systems, we hope to find what is most relevant to the user based on their previous consumption choices.

In this section, the reader will be familiarized with clustering techniques as a precursor to understanding three well-known methods used to identify the most important components of a network. In graph clustering, vertices that are closely "related" to each other are grouped into clusters. The relation of vertices depends on the context in which the graph has been produced. Going back to the airline routes example, two cities (vertices) can be thought of as related if there exists a flight that goes from any one city to the other. The cities can also be "related" in other ways such as similar flight timings or flying conditions. In general, vertices of the same cluster should be heavily connected to other vertices within the same cluster, while being sparsely connected to the rest of the graph [5]. Within graph clustering, we dominantly have global clustering, local clustering, and spectral clustering algorithms. Global methods can further be divided into iterative, divisive, and agglomerative methods.

1. **Global Techniques**

   (a) Iterative methods generally go through each vertex in the graph and assign them to a cluster. These decisions are usually not final and each vertex can be revisited for assignment, aiming to improve the optimization process [5]. The clusters can also be gradually updated when a relevant vertex is being processed. Clusters can contain sub-clusters implying that one vertex in a cluster may also belong to a larger cluster. For example, students are a part of a university (larger cluster) but they also belong to the several clubs on campus (smaller clusters). This means that graphs can have a hierarchical structure which further helps in grouping vertices by relating vertices if they have a common ancestor cluster [5].

Figure 1.1: A Hierarchical Structure



(b) Divisive techniques make use of the hierarchical structure of graphs. From the original graph, the method recursively splits clusters off in a top-down fashion [5].

(c) Agglomerative methods can substitute divisive methods. Here, algorithms start with an empty cluster and then add new clusters or assign processed vertices to existing clusters. An important point in this method is that the order in which vertices are presented to the algorithm can significantly change the clustering output. Therefore, having the opportunity to revisit vertices is essential. Irreversible clustering decisions can lead to sub-optimal results, therefore most clustering techniques are able to revise their decisions in later iterations [5].

2. **Local Techniques** A local clustering technique usually finds a solution containing or near a given vertex without looking at the whole graph. Such an algorithm operates by first selecting a seed vertex and then the algorithm only operates and assigns clusters to those vertices that surround the seed vertex up to a small maximum distance. Local algorithms can help lead to global clusters if we run a local algorithm on each vertex and combine the results to determine a global cluster [5].

1.3 Instagram: the photo-video sharing app

Having addressed the almost omnipresent nature of recommendation systems, from here on this thesis will focus on the social media giant Instagram. Founded in 2010, Instagram is a photo-video sharing application that has gained immense popularity since its origin. As of 2022, Instagram has close to 1.3 billion registered users, making it one of the most frequented social media platforms. In order to better understand the experiment conducted by the author of this work, it is necessary to understand how the platform functions.

A first-time user, USER A, needs to create an account. The account can be private or public. In the former case, any other user interested in seeing the shared content must send USER A a "follow request" and wait for their permission to be followed. In the latter, USER A can be instantly followed by anyone on the platform. Naturally, accounts related to marketing or users promoting any kind of product (or themselves) are usually public accounts and have many "followers". On USER A's daily feed, there will be posts from accounts they follow. Users can "like" posts that appear on their daily feed as well as post pictures from their accounts. Moreover, Instagram allows users to "tag" other people appearing in their posts as well as tag other accounts that may be related to their posts.

It is clear that the app has various interactive features. However, two of the most used segments are the "Explore" page and the "Reels" tab. The Explore page is tailored to give users the ability to discover new content that might not usually appear on their daily feed. The Reels tab, similar to TikTok, gives access to entertaining short videos.

Although people often inaccurately assume that there is a uniform "Instagram Algorithm" that suggests new material to users, in reality, the content suggested to users in each Instagram media component is managed by a distinct recommendation system. The experiment at the core of this work will enable us to rank the preferences of a group of Instagram users. This will be done by applying different mathematical techniques on the data collected by tracking the activity of the participants on Instagram. With the preferences obtained as numerical results from the experiment, we will further interview the participants to see whether they accept or reject what we provide them with as recommendations.

Chapter 2

Graph Ranking: Theory and Algorithms

2.1 Preliminaries

Linear algebra plays an essential role in graph analysis. In this section we provide the reader with basic terminology and background that will be used throughout this work.

2.1.1 The Eigenvalue Problem

Suppose that A is an $n \times n$ matrix. A non-zero vector $\mathbf{v}$ that satisfies the equation $A\mathbf{v} = \lambda\mathbf{v}$ is called an eigenvector of the matrix A. The scalar λ is the eigenvalue associated with this eigenvector.

In other words, if A is a linear operator, the action of A on $\mathbf{v}$, affects its length but not its direction.

In order to find the eigenpairs $\{\lambda_i, \mathbf{v}_i\}$ with $i = 1, ..., n$, we need to find the roots of the characteristic polynomial of A. That is we need to solve the problem

$$det(A - \lambda I) = 0 \tag{2.1}$$

where $det(\cdot)$ denotes the *determinant* of matrix A and I is the identity matrix. Once we have the eigenvalues, in order to find the eigenvectors, we need to solve the linear systems

$$(A - \lambda_i I)\mathbf{v}_i = 0 \quad \text{with } i = 1, ..., n$$

Remark. : The matrix $A - \lambda I$ is singular, therefore the system has infinitely many solutions.

Let $A \in \mathbb{R}^{n \times n}$ be a symmetric matrix, i.e., $A = A^T$, where A^T denotes the transpose of A. Then the *eigendecomposition* of A can be written as

$$A = V \Lambda V^T \tag{2.2}$$

where Λ is the $n \times n$ diagonal matrix whose diagonal entries are the distinct real eigenvalues of matrix A. The matrix $V = [\mathbf{v_1}, \mathbf{v_2}, \cdots, \mathbf{v_n}]$ is an *orthogonal* matrix, implying that $V^T V = V V^T = I$, that contains as columns the n eigenvectors of A. Although we required a *symmetric matrix* for the definition of eigendecomposition above, often we will need to work with *non-symmetric* matrices. The eigenvalues of a non-symmetric matrix can be complex, therefore it may be preferable to use the singular values of a matrix that are computed using the *singular value decomposition* [3].

2.1.2 Singular Value Decomposition

Singular value decomposition (SVD) is a factorization of a real, or complex, matrix into three matrices. Unlike eigendecomposition, SVD can be applied to a rectangular, non-symmetric matrix and hence is more useful for data collected from real-world scenarios.

The SVD of an $n \times m$ matrix A is given by

$$A = U\Sigma V^T \tag{2.3}$$

In Equation (2.3), U and V are orthogonal matrices, implying that $U^TU = UU^T = I$ and $V^TV = VV^T = I$. The columns of U and V contain the left and right *singular vectors* of A respectively. The diagonal entries, $d_1, d_2 \cdots, d_{min(n,m)}$, of the $n \times m$ diagonal matrix Σ represent the *singular values* of A such that

$$d_1 \geq d_2 \geq \cdots \geq d_{min(n,m)} \tag{2.4}$$

The singular values of A are denoted by σ_i, where $i = 1, 2, \cdots, min(n, m)$. If a matrix A is symmetric and *positive definite*, that is for every non-zero vector $\mathbf{x}$, the scalar $\mathbf{x^T}\ A\ \mathbf{x} > 0$, then the SVD of A is the same as its eigendecomposition.

2.2 Graph Theory

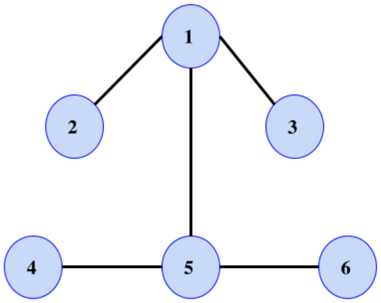
2.2.1 Basic Concepts

In mathematics, graphs are structures used to model pairwise relations between objects. A graph consists of vertices (also called nodes) connected by edges (also called links). Here are some useful definitions:

Definition 1. A graph $G = (V, E)$ is an ordered pair of a set of vertices $V = \{v_i\}$ and a set of edges $E \subseteq V \times V$.

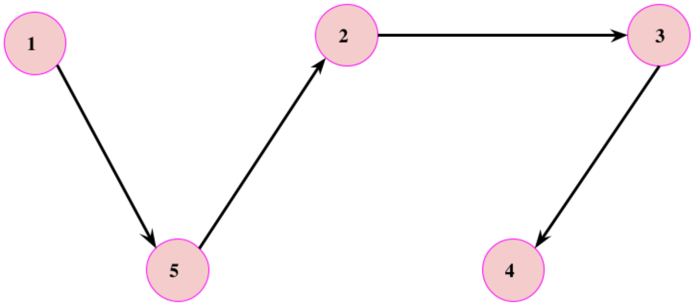
Definition 2. A graph is called *undirected* if $(i, j) \in E$ implies that $(j, i) \in E$, with $i, j = 1, 2, \cdots, n$.

Figure 2.1: An Undirected Graph



Definition 3. A *graph* is called directed if $(i, j) \in E$ does not imply that $(j, i) \in E$, with $i, j = 1, 2, \cdots, n$.

Figure 2.2: A Directed Graph



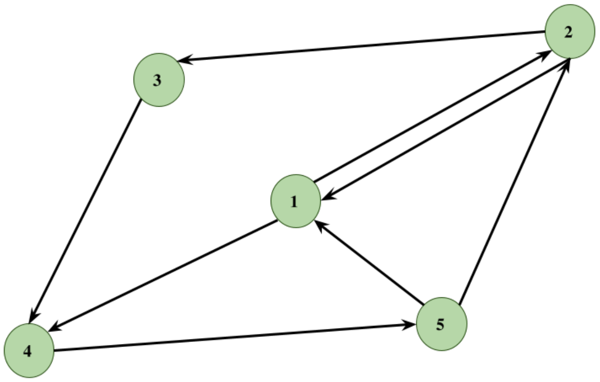
Directed graphs, also known as digraphs, are composed of edges that have arrows. In contrast, an undirected graph has bidirectional edges (with no arrows). In order to be clear about the difference, let us comment on the interactions between nodes 1 and 5 in both graphs above. In Figure 2.1 there is an edge between 1 and 5 with no arrow. This means “1 goes into 5 and 5 goes into 1”. On the contrary, in Figure 2.2, 1 and 5 are connected through an edge with a specified direction, that is, “1 goes into 5, but 5 does not go into 1”. To understand the difference between these two types of graphs in the real world, and to start building a connection with Instagram, let us think of the difference between a directed and undirected edge between two users, where the connection indicates the action of one user following another. If USER A follows USER B but USER B does not follow USER A, we would have two nodes, A and B, connected by an arrow that goes from A to B, but not from B to A. If USER A and USER B follow each other, then it is enough to draw a segment from node A to node B. It stands for a reciprocal connection between A and B.

So far, we have talked about the connection between two nodes. However, when dealing with a graph, we may want to see if some nodes are connected with others even in the absence of a direct edge.

A *path* is a finite or infinite sequence of edges which joins a sequence of vertices that can be repeated. A *walk* is a sequence of directed edges $i \longrightarrow i_1 \longrightarrow i_2 \cdots \longrightarrow i_k \longrightarrow j$ that can be repeated. If none of the vertices $i_1, i_2, \cdots, i_k$ are repeated, then the walk is called a *directed path.*

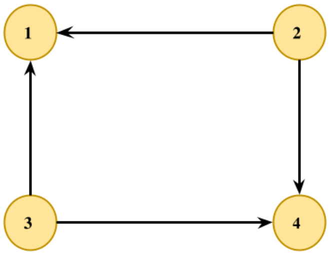
Definition 4. A directed graph is *strongly connected* if for any pair of vertices, $(i, j) \in V$, there exits a directed path from vertex i to vertex j.

Figure 2.3: Strongly Connected Digraph



Definition 5. A directed graph is *weakly connected* if replacing all of its directed edges with undirected edges produces a connected (undirected) graph.

Figure 2.4: Weakly Connected Digraph



We now require a mathematical representation of a graph.

Definition 6. The adjacency matrix $A \in \mathbb{R}^{n,n}$ of a graph $G = (V, E)$ is a square matrix whose entries are defined as follows

$$A(i,j) = a_{i,j} = \begin{cases} 1 & \text{if } (i,j) \in E \\ 0 & \text{if } (i,j) \notin E \end{cases} \tag{2.5}$$

According to the definition above, A can have only entries 0 and 1. Specifically, if there is an edge connecting two nodes i and j, then the corresponding entry, that is $a_{i,j}$ takes the value 1: it will be 0 otherwise. If the graph is undirected, the adjacency matrix A is symmetric, that is, $a_{i,j} = a_{j,i}$ for all $i, j \in V$. A is non-symmetric for a digraph.

Due to the nature of the experiment conducted, in this thesis we will mostly deal with digraphs, and therefore non-symmetric adjacency matrices. As an exercise, let us write down the adjacency matrix corresponding to the digraph shown in Figure 2.2.

$$\mathbf{A} = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix} \tag{2.6}$$

A basic property associated with vertices in a graph is their *degree.* For an undirected graph, each node has a degree d if it has d incident edges. In other words we are counting the edges that are attached to a certain node. However, for a directed graph, we must distinguish between the number of edges entering a node, the *in-degree*, and the number of edges departing from a node, the *out-degree.* Therefore, it is necessary to define two different matrices representing the degree of each node.

Definition 7. The in-degree of a node in a digraph, G, is the number of directed edges coming into the node. The out-degree of a node in G is the number of directed edges starting at (or coming out of) the node. Let us denote the corresponding in-degree and out-degree matrices by

$$D_{in} = diag(d_1^{in}, d_2^{in}, \ldots, d_n^{in}) \quad \text{and} \quad D_{out} = diag(d_1^{out}, d_2^{out}, \ldots, d_n^{out})$$

Remark. We can observe that the row sums of the adjacency matrix A correspond to the out-degrees of G, that is,

$$(A\mathbb{1})_i = d_i^{out} \quad \text{for } i = 1, \ldots, n \tag{2.7}$$

The column sums of the adjacency matrix A correspond to the in-degrees of G, that is,

$$(A^T\mathbb{1})_i = d_i^{in} \quad \text{for } i = 1, \ldots, n \tag{2.8}$$

Let us consider again the graph in Figure 2.2. In this case,

$$D_{in} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \quad D_{out} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{2.9}$$

We have previously introduced strongly and weakly connected digraphs. When it becomes necessary to classify G as strongly or weakly connected, we can directly verify some properties of the adjacency matrix. In particular, one can exploit the *reducibility* of the adjacency matrix A.

Definition 8. A square matrix A is said to be reducible if there exists a permutation matrix Π such that

$$\Pi\ A\ \Pi^T = \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix}$$

where A_{11} and A_{22} are both square (that is, A is equivalent to a block triangular structure). If there is no such Π, we say that A is irreducible.

Theorem 2.2.1. *A digraph is strongly connected if and only if the corresponding* *adjacency matrix A is irreducible.*

Remark. For any diagraph $G = (V, E)$ there exists a permutation matrix Π such that

$$\Pi\ A\ \Pi^T = \begin{bmatrix} A_{11} & A_{12} & \cdots & \cdots & A_{1p} \\ 0 & A_{22} & \cdots & \cdots & A_{2p} \\ \vdots & \vdots & \ddots & & \vdots \\ \vdots & \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \cdots & A_{pp} \end{bmatrix}$$

where each diagonal block is square and irreducible.

The subgraphs of G having the adjacency matrices $A_{11}, A_{22}, \ldots, A_{pp}$ are called strongly connected components of G. In particular, the largest one is called the *maximal strongly connected component.*

2.2.2 Graph Laplacians

Graph Laplacians are another way to represent graphs, and they are particularly useful as a bridge between the discrete representations of a graph and continuous representations such as vector spaces. In particular, they play a fundamental role in clustering.

For an undirected graph G, the Laplacian is defined as

$$L = D - A, \tag{2.10}$$

where A is the adjacency matrix of G and D is its *degree matrix.* As for an undirected graph, the in-degree of each node is equal to its out-degree, hence there is only one degree matrix D of G.

How do we define the Laplacian for a directed graph if we have two different degree matrices to consider? While there is a unique definition for the Laplacian of an undirected graph, there is no such counter-part for the directed case. In the next section we introduce two ways of defining and understanding the Laplacian of a directed graph.

2.2.3 Laplacians for Directed Graphs

Here we describe two ways to define a Laplacian for digraphs:

- Turn a digraph into an undirected graph, by making use of a bi-partition of G.
- Renounce the symmetry and define two distinct Laplacians: the *in-Laplacian* and the *out-Laplacian*.

1. **Symmetric Laplacians via a Bipartite Model**

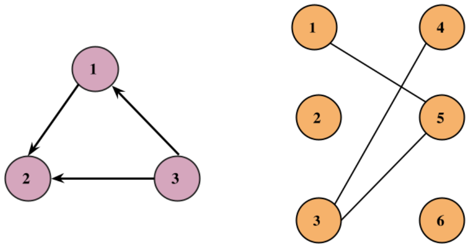
   Definition 9. An undirected graph $\mathcal{V} = (\mathcal{V}, \mathcal{E})$ is bipartite if $\mathcal{V} = \mathcal{V}_1 \cup \mathcal{V}_2$ with $\mathcal{V}_1 \cap \mathcal{V}_2 = \emptyset$.

   In the definition above, $\mathcal{V}_1$ and $\mathcal{V}_2$ are sets of nodes such that nodes in $\mathcal{V}_1$ can only be connected to nodes in $\mathcal{V}_2$, and nodes in $\mathcal{V}_2$ can only be connected to nodes in $\mathcal{V}_1$.

   Remark. Any digraph on n nodes can be uniquely represented by a bipartite graph on $2n$ nodes as follows: let $G = (V, E)$ be the digraph, a bipartite graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ can be constructed by defining $\mathcal{V} = V \cup V'$, where $V = \{1, 2, \ldots, n\}$, $V' = \{n+1, n+2, \ldots, 2n\}$ and $\mathcal{E} = \{(i, j') \mid j' = n + j,\ (i, j) \in E\}$.

   The figure shown below illustrates this procedure.

Figure 2.5: A Digraph and its Bipartite Graph



The adjacency matrix for the digraph shown above is,

$$A = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 1 & 1 & 0 \end{bmatrix} \tag{2.11}$$

It is not hard to recognize that the adjacency matrix of the bipartite graph can be written as

$$\mathcal{A} = \begin{bmatrix} 0 & A \\ A^T & 0 \end{bmatrix} \tag{2.12}$$

Thus, it is possible to define the Laplacian as $\mathcal{L} = \mathcal{D} - \mathcal{A}$ where $\mathcal{D}$ is the degree matrix relative to the bipartite graph. The Laplacian, in terms of the original adjacency matrix A and the out-degree and in-degree matrices of our original digraph can be written as

$$\mathcal{L} = \begin{bmatrix} D_{out} & -A \\ -A^T & D_{in} \end{bmatrix} \tag{2.13}$$

Notice that $\mathcal{L}$ is symmetric. Therefore, using a bipartite graph, we obtain a symmetric Laplacian for our original digraph.

2. **Non-symmetric Laplacians**

We now define two distinct Laplacians: one per each degree matrix of the digraph G.

$$L_{in} = D_{in} - A \quad \text{and} \quad L_{out} = D_{out} - A. \tag{2.14}$$

Here, A represents the adjacency matrix of G, D_{in} is the in-degree matrix of G, and D_{out} is the out-degree matrix of G. Since L_{in} and L_{out} are non-symmetric, as stated previously, their eigenvalues will be complex. Consequently, we cannot associate a physical phenomenon to this.

Before moving further with the concept of non-symmetric Laplacians, it is important to have a general understanding of the Perron-Frobenious theorem. The study of the asymptotic behavior of matrices with non-negative entries is the essence of Perron-Frobenious theory. The theorem explores the properties of the *spectral radius* of non-negative matrices ($A > 0$). The spectral radius of a square matrix A is the largest absolute value of its eigenvalues and it is often denoted as $\rho(A)$.

Theorem 2.2.2. *Perron 1903. Let* $A \in \mathbb{R}^{n\times n}, A > 0$. *Then:*

(a) $\rho(A) > 1$

(b) $\rho(A)$ *is an eigenvalue of* A *and it is simple (i.e., it has algebraic multiplicity one)*

(c) $\exists x \in \mathbb{R}^n, x > 0$ *such that* $Ax = \rho(A)x$

(d) $\rho(A)$ *is the only eigenvalue of largest modulus*

Coming back to our discussion on non-symmetric Laplacians, let us assume that $G = (V, E)$ is strongly connected. Then the adjacency matrix A is irreducible and non-negative. By the Perron-Frobenius theorem, the spectral radius of A is a simple eigenvalue of A and its associated eigenvector has all positive components. We will introduce in the section addressing PageRank that this eigenvector, with all positive components, gives us the weights for the importance of the nodes in G.

As a consequence of G being strongly connected, 0 is a simple eigenvalue of L_{in} and L_{out}. Further, since G is strongly connected, D_{in} and D_{out} are invertible since strong connectivity implies nonzero diagonal entries for the degree matrices. Therefore,

$$L_{out} = (I - AD_{out}^{-1})D_{out} \quad \text{and} \quad L_{in} = D_{in}(I - D_{in}^{-1}A) \tag{2.15}$$

Now, $A \geq 0$, $D_{in}^{-1} \geq 0$ and $D_{out}^{-1} \geq 0$ imply that AD_{out}^{-1} and $D_{in}^{-1}A$ are both non-negative and irreducible. Hence, by the Perron-Frobenius theorem, both matrices have a simple dominant eigenvalue of $\lambda = 1$.

2.3 Ranking Methods Used in Our Experiment

There are various methods used to compute vertex (node) prestige, but a few of them use the notion of nodes as *hubs* and *authorities*. In a network, a node is a hub if it broadcasts information while an authority is a node that receives information [1]. Note, a node can be a hub and an authority. In this section we introduce three well-known ranking methods, namely, Dominant Eigenvector, HITS, and PageRank.

2.3.1 Dominant Eigenvector

The Dominant Eigenvector method computes eigenvector centrality and prestige to identify the most important hub or authority in a network. This approach takes into account an intuitive way of recognizing an important node. A node is considered important if it interacts with other important nodes, which in turn increase its prestige. Using the example of an adjacency matrix, containing zeros and ones that mean "do not like" and "like" respectively, the first fundamental step toward the dominant eigenvector approach was made by Seely [6]. He noted that it is important to be liked by someone who is in turn being liked a lot by others. In other words, a node's index of prestige should account for the prestige of the nodes that endorse it.

The idea proposed by Seely, denoted by the function $r(\cdot)$, can be formalized as follows

$$r(v) = \sum_{u \in V} A(u, v) r(u). \tag{2.16}$$

Equation (2.16) corresponds to the set of $|V|$ linear equations and can be rewritten as

$$\mathbf{r} = \mathbf{A^T r}, \tag{2.17}$$

where vector $\mathbf{r}$, of size $|V|$, stores all rank scores, and $\mathbf{A}$ is the adjacency matrix of the network [5].

In order for Equation (2.17) to have a finite solution Katz proposed that matrix $\mathbf{A}$ be manipulated so that every row in $\mathbf{A}$ has a sum equal to 1 [5]. Then, from Equation (2.17), we can conclude that $\mathbf{r}$ is an eigenvector of $\mathbf{A^T}$ with a corresponding eigenvalue of 1. Consequently, another form of Equation (2.17) was suggested by Bonacich [2] where it is assumed that the rank of each vertex is proportional to the weighted sum of the vertices it is connected to. This conceptualization results in the dominant eigenvector approach, also known as *eigenvector centrality* and can be expressed as

$$\lambda \mathbf{r} = \mathbf{A^T r} \tag{2.18}$$

Given (2.18), the hub and authority measures for a digraph are calculated as follows. With the adjacency matrix A of the digraph we compute

$$A\mathbf{x} = \lambda \mathbf{x} \tag{2.19}$$

and

$$A^T\mathbf{y} = \lambda\mathbf{y} \tag{2.20}$$

Using Equation (2.19), we compute the eigenvector of A corresponding to its largest eigenvalue, hence called the dominant eigenvector [1]. The dominant vector then contains the hubs scores for each node in the graph. Similarly, using Equation (2.20), we compute the dominant eigenvector of A^T which contains the authority scores for each node in the graph.

2.3.2 HITS

Hyperlink Induced Topic Search (HITS) is a method used for link analysis. Originating from the idea of ranking and discovering web pages based on a particular search, HITS helps us in finding relevant information based on our search requests. The idea of the method is motivated by the fact that an ideal website should link to other relevant sites and should also be linked to by other important sites.

HITS computes two vectors, $\mathbf{u_1}$ and $\mathbf{v_1}$, representing the scores of the nodes as broadcasters or receivers respectively. The nodes corresponding to the maximum scores of $\mathbf{u_1}$ and $\mathbf{v_1}$ will be the most important hub and authority. In order to compute the ranking vectors, we can use the SVD of the adjacency matrix of a digraph as the digraphs, and adjacency matrices, used in the experiment discussed later involve a small data set. For larger data sets, it is important to note that using the SVD is computationally expensive and an iterative algorithm of HITS is better suited in that case. Let us see how all of this relates to the mathematical tools we discussed.

Let A be the adjacency matrix of the digraph G. HITS performs a Singular Value Decomposition (SVD) of A, that is, $A = U\Sigma V^T$. Therefore the SVD of $A^T = (U\Sigma V^T)^T = V\Sigma U^T$. Here $U = [\mathbf{u_1} \ \ \mathbf{u_2}....\mathbf{u_n}]$ is the matrix of left singular vectors of the adjacency matrix A while $V = [\mathbf{v_1} \ \ \mathbf{v_2}....\mathbf{v_n}]$ is the matrix of the right singular vectors of A. The matrix $\Sigma = diag(\sigma_1, \sigma_2, \cdots, \sigma_n)$ is the matrix of the singular values of A.

Let us define the hub matrix as $A_1 = AA^T$ and the authority matrix as $A_2 = A^T A$. While A is not symmetric as it corresponds to a digraph, A_1 and A_2 are now symmetric. Using the SVD of A we have

$$A_1 = AA^T = (U\Sigma V^T)(V\Sigma U^T) = U\Sigma^2 U^T \tag{2.21}$$

$$A_2 = A^T A = (V\Sigma U^T)(U\Sigma V^T) = V\Sigma^2 V^T \tag{2.22}$$

Since we are interested in using the hub and authority matrices we come to the following conclusions. Given that $A = U\Sigma V^T$, we can compute that $AV = U\Sigma$. Using the left singular values, $A\mathbf{v_i} = \sigma_\mathbf{i}\mathbf{u_i}$ [1]. Similarly for $A^T = V\Sigma U^T$, $A^T U = \Sigma V$. Now, using the right singular values, we compute that $A^T\mathbf{u_i} = \sigma_\mathbf{i}\mathbf{v_i}$. Since we are interested in using the hub and authority matrices we come to the following conclusions.

$$\begin{aligned}
A^T\mathbf{u_i} &= \sigma_i\mathbf{v_i} \\
AA^T\mathbf{u_i} &= A_1\mathbf{u_i} \\
&= A(\sigma_i\mathbf{u_i}) \\
&= \sigma_i(A\mathbf{u_i}) \\
&= \sigma_i(\sigma_i\mathbf{u_i}) \\
A_1\mathbf{u_i} &= \sigma_i^2\mathbf{u_i}.
\end{aligned} \tag{2.23}$$

Similarly,

$$\begin{aligned}
A\mathbf{v_i} &= \sigma_i\mathbf{u_i} \\
A^TA\mathbf{v_i} &= A_2\mathbf{v_i} \\
&= A^T(\sigma_i\mathbf{u_i}) \\
&= \sigma_i(A^T\mathbf{u_i}) \\
&= \sigma_i(\sigma_i\mathbf{v_i}) \\
A_2\mathbf{v_i} &= \sigma_i^2\mathbf{v_i}.
\end{aligned} \tag{2.24}$$

Given the results of (2.23) and (2.24) respectively, we can conclude that $\mathbf{u_i}$ is the i^{th} left eigenvector of the hub matrix $A_1 = AA^T$ and $\mathbf{v_i}$ is the i^{th} right eigenvector of the authority matrix $A_2 = A^TA$. Assume $i = 1$, then $\mathbf{u_1}$ in our HITS algorithm contains the hub scores for each node of the directed graph, and $\mathbf{v_1}$ contains the authority scores for each node.

2.3.3 PageRank

PageRank is an algorithm used by Google Search to rank web pages. The PageRank metric is a representation of how important Google thinks a particular web page is relative to the search query. For instance, a score of 0 is associated with low-quality websites while a score of 10 would represent the most authoritative pages on the web. Now that we are familiar with non-symmetric Laplacians for directed graphs, we can understand the PageRank method that was used during the experiment conducted for this thesis. However, we need further mathematical definitions to be able to illustrate the method.

1. **Stochastic Matrices** A stochastic matrix is a square matrix whose columns are *probability vectors*. A probability vector is a vector whose entries are real numbers between 0 and 1 and whose sum is 1.

2. **Markov Chain** In a Markov chain, elements move from one state to another with the same probabilities at each step in the process.

3. **Transition Matrix** The transition matrix for a Markov chain is a stochastic matrix whose (i, j) entry gives the probability that an element moves from the j-th state to the i-th state during the next step of the process.

We can now proceed with describing the PageRank method. We first normalize our out-Laplacian. Thus we define a matrix H such that

$$H := I - A^T D_{out}^{-1} = L_{out}^T D_{out}^{-1} \tag{2.25}$$

Here, H is a column stochastic matrix, its zero eigenvalue is simple and the solution of $H\mathbf{x} = \mathbf{0}$ is the stationary probability distribution of the Markov chain described by the transition matrix $A^T D_{out}^{-1}$. The vector $\mathbf{x}$ in this case is also known as the PageRank vector with all positive components that give us the scores for the importance of the nodes in digraph G. The hub scores for each node are stored in vector $\mathbf{x}$.

Remark. If there are nodes in the digraph with a zero out-degree, the corresponding matrix H will not be stochastic. In such cases the adjacency matrix needs to be modified. Through the experiment conducted in this thesis, we came across non-stochastic adjacency matrices. In order to convert them to stochastic ones, we replaced each row entry of the node with a zero out-degree with $\frac{1}{n}$ in the adjacency matrix, where n is the number or nodes.

PageRank does not distinguish between hub and authority rankings, but we can compute a *reverse* PageRank vector. To this end, we normalize the in-Laplacian. Thus we define a matrix K such that

$$K := I - AD_{in}^{-1} = L_{in} D_{in}^{-1} \tag{2.26}$$

In this case as well, there exists a unique probability distribution $\mathbf{y}$ that satisfies $K\mathbf{y} = \mathbf{0}$, which is the probability distribution of the Markov chain described by the column stochastic matrix AD_{in}^{-1}. The reverse PageRank vector $\mathbf{y}$ contains the authority scores for each node in G.

Remark. The normalized matrices H and K can be also written in terms of the reversed digraph G', which is obtained from G by reversing the directions on the edges. Given the adjacency matrix A of G, $A' = A^T$ is the adjacency matrix of G'. Moreover $D'_{out} = D_{in}$ and $D'_{in} = D_{out}$.

Therefore,

$$L'_{out} = D'_{out} - A' = D_{in} - A^T = L_{in}^T \text{ and } L'_{in} = D'_{in} - A' = D_{out} - A^T = L_{out}^T \quad (2.27)$$

Thus

$$H = L_{out}^T D_{out}^{-1} = L'_{in}(D'_{in})^{-1} \text{ and } K = L_{in}^T D_{in}^{-1} = L'_{out}(D'_{out})^{-1} \quad (2.28)$$

Chapter 3

An Instagram Experiment

One of the main goals of this project was to draw connections between social media used by the author everyday and her favorite field of mathematics: linear algebra. To make this experiment personal, we decided to involve Emory students and create our data set, rather than referring to an available database. Therefore, the data collection process, and its interpretation, has been an exciting and delicate procedure.

3.1 Data Collection

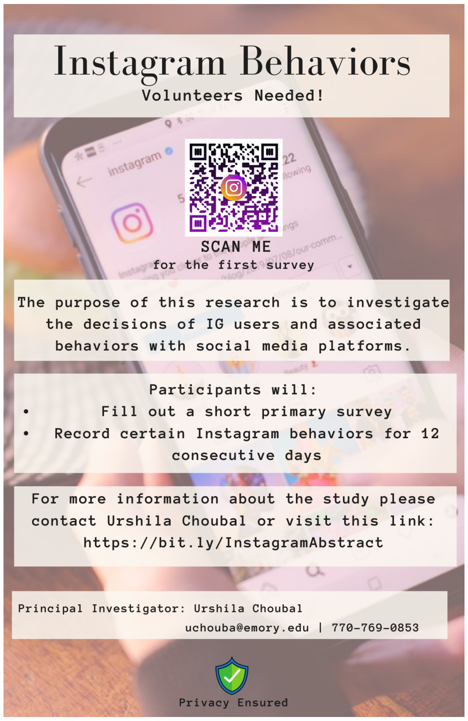
The primary goal behind the data collection process was to create a set of directed graphs. They were produced by keeping track of the users' activity on Instagram. A secondary goal of the experiment was to figure out what Instagram content was popular among users and what content was unpopular. The experiment then took into account the topics of disinterest to set up a stage to facilitate the primary goal. To this end, we created an Instagram account and asked a group of Emory students to visit the account for 12 consecutive days. We aimed to keep a track of their habits and preferences to obtain data that could be mathematically represented as directed graphs. Once the graphs were created, we were able to analyze the data through the methods we discussed in Chapter 2, that is, HITS, Dominant Eigenvector, and

PageRank. The detailed steps of the data collection process have been outlined below.

1. **Gathering participants and the primary survey**

   With the goal of creating a set of directed graphs, the first step was to scan the Instagram consumption habits of different users. In order to mobilize participants for the experiment, we created the following poster.

Figure 3.1: Poster Used to Publicize the Experiment



Instagram Behaviors

Volunteers Needed!

SCAN ME

for the first survey

The purpose of this research is to investigate the decisions of IG users and associated behaviors with social media platforms.

Participants will:
- Fill out a short primary survey
- Record certain Instagram behaviors for 12 consecutive days

For more information about the study please contact Urshila Choubal or visit this link: https://bit.ly/InstagramAbstract

Principal Investigator: Urshila Choubal
uchouba@emory.edu | 770-769-0853

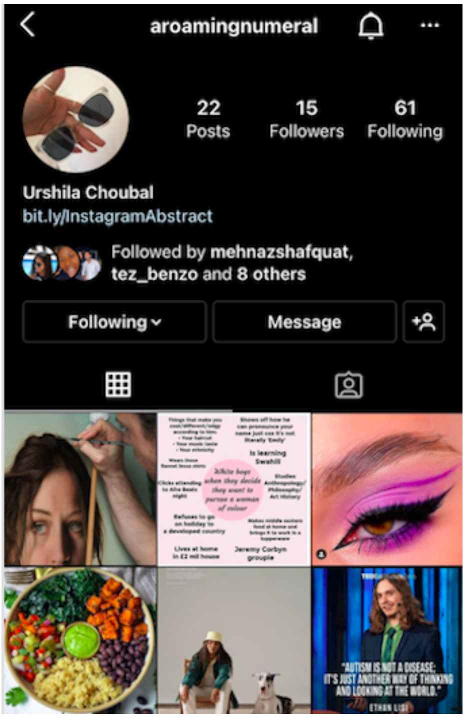
Privacy Ensured

The poster was widely broadcasted among various Emory communities, and interested participants were instantly navigated to a primary survey titled "Instagram Behaviors". Besides wanting to gain insight into the general activity of users on Instagram, we primarily wanted to know what content was of importance to the Emory community. We were also curious to know what type of content did not interest the participants. From a list of twenty categories of content, the survey asked participants to rank each category from one to six, one being the most preferred, and six being the least preferred category. Out of the fifty-five individuals that took the primary survey, the unpopular content topics were the following eight: safety, home/culture/lifestyle, companies and institutions, climate and animal justice, sports, social justice, education, beauty/clothing/self care, and art. For the rest of the experiment, only nine out of the fifty-five initial participants were included.

2. **Setting up the mock Instagram account**

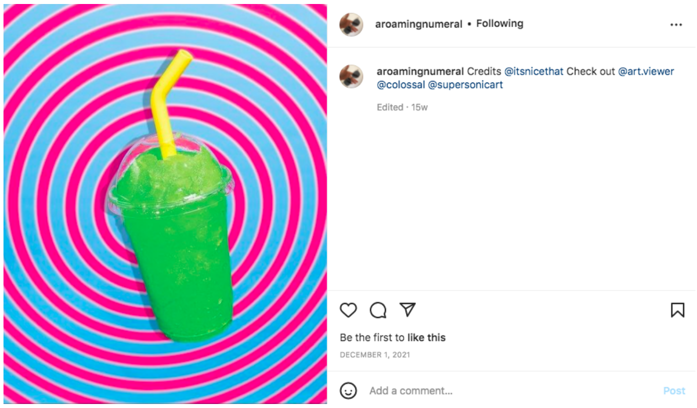
   The next step of the process involved tracking the activity of the nine core participants on Instagram on a daily basis. In order to expose the participants to the content they did not have an interest in, we created a mock Instagram account named "A Roaming Numeral" which only included posts related to the aforementioned eight topics.

Figure 3.2: Experimental Instagram Account: @aroamingnumeral


aroamingnumeral
22 Posts
15 Followers
61 Following
Urshila Choubal
bit.ly/InstagramAbstract
Followed by mehnazshafquat, tez_benzo and 8 others
Following
Message
"AUTISM IS NOT A DISEASE; IT'S JUST ANOTHER WAY OF THINKING AND LOOKING AT THE WORLD."
ETHAN LISI

Over the span of twelve days various pictures were shared through the account, and participants had to engage with them on a daily basis. Using any picture, from our main account, participants were instructed to further move onto different accounts from those that were tagged in our picture. The accounts that were tagged on any post from our account were always related by content to that post. For example, consider the post of the green drink below.

Figure 3.3: An Example of a Post from the Mock Account



aroamingnumeral • Following

aroamingnumeral Credits @itsnicethat Check out @art.viewer @colossal @supersonicart

Edited · 15w

Be the first to like this

DECEMBER 1, 2021

The picture was originally posted by the account @itsnicethat, and in the caption we can see that three other accounts have been tagged namely @art.viewer @colossal and @supersonicart. These three accounts are related to the account @itsnicethat because all three promote creativity and art.

3. **Participant engagement with the account**

   It is important to understand how a participant interacted with the mock account, and how that information was later used to create a directed graph specific to that user. Assume a participant chose the green drink post, credited to the account @itsnicethat, as their first post to interact with. The participant could then choose to move onto any of the accounts we tagged in our post. Assume they choose to have look at the account @supersonicart. Now the participant is free to browse this account, however, they must move onto a third account so that we have enough data to reach a conclusion. Choosing the third account (i.e., @thewalusogallery) had no restrictions except that it had to be from the list of people the account @supersonicart followed. From here on the participant is free to browse this third account, continue jumping to different accounts, or terminate the entire process. As such the “pathway” for this set of decisions, for this participant, would look like

$$@itsnicethat \longrightarrow @supersonicart \longrightarrow @thewalusogallery.$$

Therefore, each pathway had to have a minimum of three accounts, and the participant was instructed to give us one pathway per day. A participant could provide us with more than one pathway. However, in order to start a new pathway, they had to restart the decision-making process by choosing another basis post from our mock account. Participants usually provided us with two to three pathways every day.

4. **Data collection on Excel**

   For the twelve days that the experiment was conducted, an excel sheet was updated to keep note of all the pathways taken for each of the nine core participants. The only thing left to do was turn this data into directed graphs. In order to create the graphs, we decided to analyze the content of the pathways. Coming back to the example pathway

$$@itsnicethat \longrightarrow @supersonicart \longrightarrow @thewalusogallery,$$

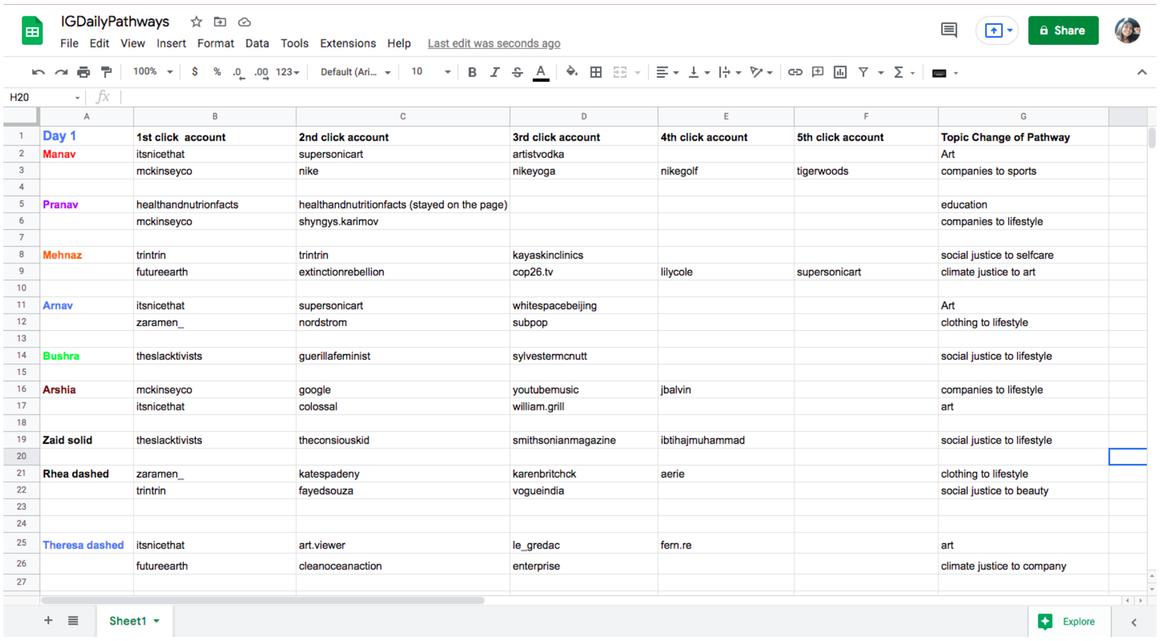
   we can see that the user initially started with an art account but ended up at an account (@thewalusogallery) that has been enlisted under the “company” title on Instagram since it is an art gallery. Therefore in reference to the eight topics of disinterest mentioned earlier, the participant moved from the Art category to the Companies and Institutions one. As an additional example, consider the pathway

$$@trintrin \longrightarrow @fayedsouza \longrightarrow @vogueindia.$$

The starting account, @trintrin, posts content related to social justice, while the last account of the pathway (@vogueindia) is related to clothing and beauty. Therefore, in this case, the participant jumped from the topic of Social Justice to the topic of Beauty/Clothing/Self care.

In order to rank the user's preferences, we were interested in seeing the change in topic that occurred in the pathways as users moved from one account to the other and wanted to represent this clearly through a graph. For this reason each node in our graphs represent a topic and each edge an existing pathway.

Figure 3.4: Data Collected on Day 1


IGDailyPathways
File Edit View Insert Format Data Tools Extensions Help
Last edit was seconds ago
Share
100%
Default (Ari...
10
H20
Sheet1
Explore

Day 1	1st click account	2nd click account	3rd click account	4th click account	5th click account	Topic Change of Pathway
Manav	itsnicethat	supersonicart	artistvodka			Art
	mckinseyco	nike	nikeyoga	nikegolf	tigerwoods	companies to sports
Pranav	healthandnutrionfacts	healthandnutritionfacts (stayed on the page)				education
	mckinseyco	shyngys.karimov				companies to lifestyle
Mehnaz	trintrin	trintrin	kayaskinclinics			social justice to selfcare
	futureearth	extinctionrebellion	cop26.tv	lilycole	supersonicart	climate justice to art
Arnav	itsnicethat	supersonicart	whitespacebeijing			Art
	zaramen_	nordstrom	subpop			clothing to lifestyle
Bushra	theslacktivists	guerillafeminist	sylvestermcnutt			social justice to lifestyle
Arshia	mckinseyco	google	youtubemusic	jbalvin		companies to lifestyle
	itsnicethat	colossal	william.grill			art
Zaid solid	theslacktivists	theconsiouskid	smithsonianmagazine	ibtihajmuhammad		social justice to lifestyle
Rhea dashed	zaramen_	katespadeny	karenbritchck	aerie		clothing to lifestyle
	trintrin	fayedsouza	vogueindia			social justice to beauty
Theresa dashed	itsnicethat	art.viewer	le_gredac	fern.re		art
	futureearth	cleanoceanaction	enterprise			climate justice to company

3.2 Data Interpretation

The next step in the experimental process was to create directed graphs for each participant based on the pathways they provided us with for all twelve days. However, this was not our original intention. Our first idea was to create one singular directed graph that contained all the pathways that originated from five chosen pictures from our account. As participants were allowed to pick any initial image each day, on several occasions, various participants decided to revisit a picture they had already chosen before leading to an entirely new pathway. While the directed graph produced from these pathways could have given different conclusions, it later made more sense to focus on the change in the topic of the pathways for each participant individually.

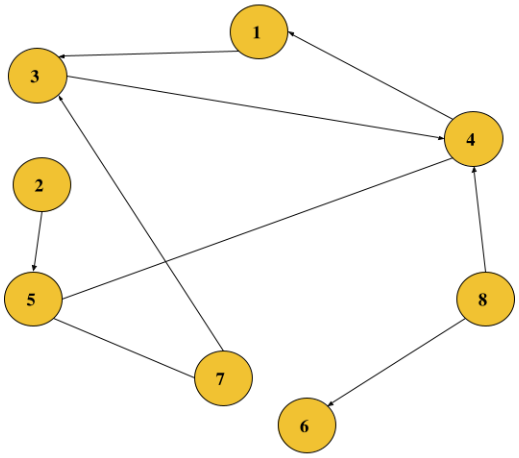
Therefore, we assigned numbers to the eight topic categories which represented the eight nodes in each directed graph. The list was as follows:

1. Art
2. Climate and Animal justice
3. Companies and Institutions
4. Social Justice
5. Home/Culture/Lifestyle
6. Sports
7. Beauty/Clothing/Self Care
8. Education

As mentioned, the edges connecting the nodes represent the change in topic along a user's pathway. Consider the example shown in Figure 3.5. The arrow going from node 4 into node 1 represents a pathway from Social Justice to Art. The line segment between between node 4 and 5 implies that there was a pathway leading from Social Justice (node 4) to Home/Culture/Lifestyle (node 5) at some instance, but the reverse also took place. That is, a participant also moved from Home/Culture/Lifestyle (node

5) to Social Justice (node 4) in some pathway. Through this process, we constructed nine distinct directed graphs corresponding to the nine participants.

Figure 3.5: A Digraph from the Experiment



As introduced in Chapter 2, to each digraph we associate an adjacency matrix. This is crucial to supplement the ranking techniques. Consider, again, the graph in Figure 3.5. The corresponding adjacency matrix is

$$\mathbf{A} = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \end{bmatrix} \tag{3.1}$$

Now that we have our adjacency matrix, we can use this as the input for the ranking methods discussed in chapter 2. We can visually make some conclusions about which nodes might be considered important in our analysis, however it is necessary to see what the rank methods determine. By importance we imply our interest in knowing which node received the highest hub score and which node received the highest authority score by the three methods. The ideal situation would be if our intuitions about the directed graph in Figure 3.5 match the results of the methods. Once the most important hub and authority nodes were determined our intention was to provide the results to the participant and give them the opportunity to either accept or reject our conclusions [4].

After having looked at the data from an individual perspective, we were motivated to group individuals into categories to see which nodes were deemed important to each group. Of the nine participants, five identified as female, while four identified as male. To create the directed graphs associated with these two categories, we constructed two new graphs. One, with all the choices (pathways) of the participants that identified as female and the other with all the choices of the participants that identified as male.

Figure 3.6: Group 1: Female

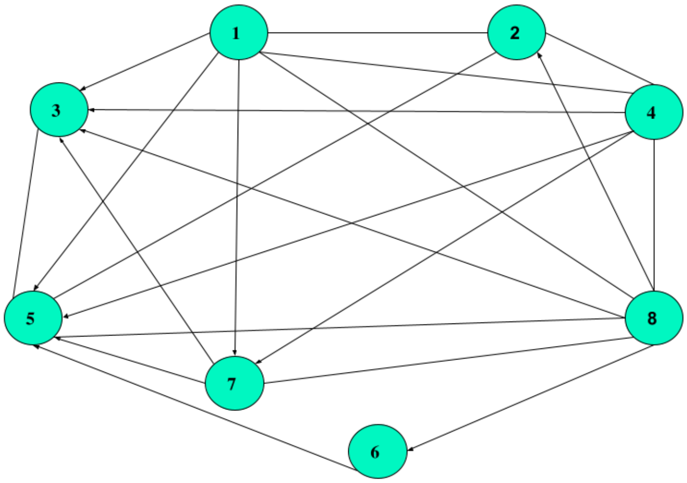
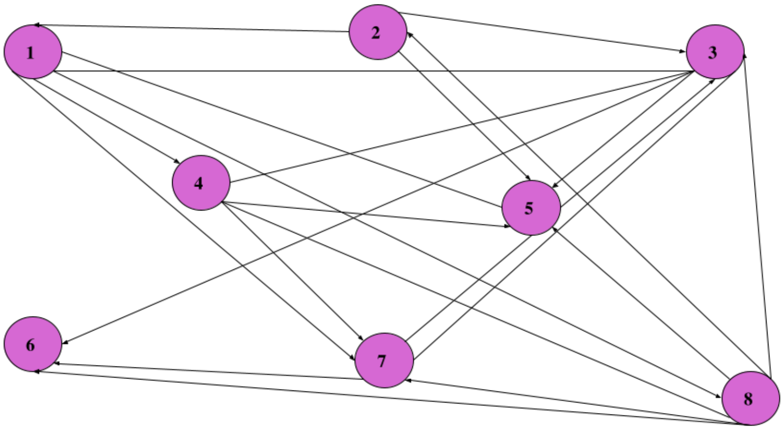


Figure 3.7: Group 2: Male



With these newly constructed graphs, we were able to compute their adjacency matrices and run the rank methods on them. The numerical results are discussed below. Specifically, in the following section we report the node ranking for each participant and for the groups (males and females). Detailed results are stored in tables in Appendix A. Each table, associated with a participant, contains the hub and authority scores of the eight nodes as computed by HITS and then Dominant Eigenvector. When the participants are grouped, we look at each group (female or male) as an individual unit and make sense of the numerical results accordingly.

3.3 Numerical Results

The numerical results from the experiment include the digraph of each participant along with an associated table containing the results. The tables below are a more concise version of the tables available in Appendix A. The tables in Appendix A are populated with the node rankings as well as the scores associated with each rank. Each table below contains the hub and authority rankings as computed using HITS and the Dominant Eigenvector approach. As previously stated, the most important hub and authority, as computed by the methods, are the first set of nodes in each column. This implies that for a specific participant, these set of nodes (therefore the respective social media category) were deemed to be important to the participant by the methods.

Figure 3.8: Directed Graph for Participant One

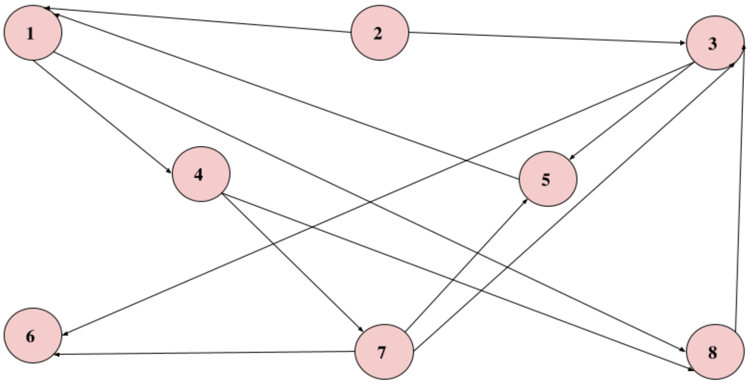


Table 3.1: Table 1

HITS		Eig	
Hub	Auth	Hub	Auth
7	3	2	6
3	5	1	5
2	6	4	3
8	1	7	8
5	2	5	1
1	4	3	4
4	7	8	7
6	8	6	2

Figure 3.9: Directed Graph for Participant Two

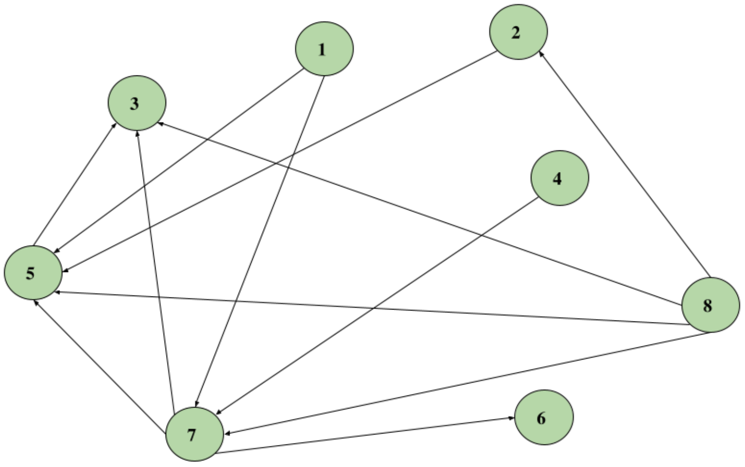


Table 3.2: Table 2

HITS		Eig	
Hub	Auth	Hub	Auth
8	5	1	3
7	3	2	1
1	7	3	2
2	2	4	4
5	6	5	5
4	1	6	6
3	4	7	7
6	8	8	8

Figure 3.10: Directed Graph for Participant Three

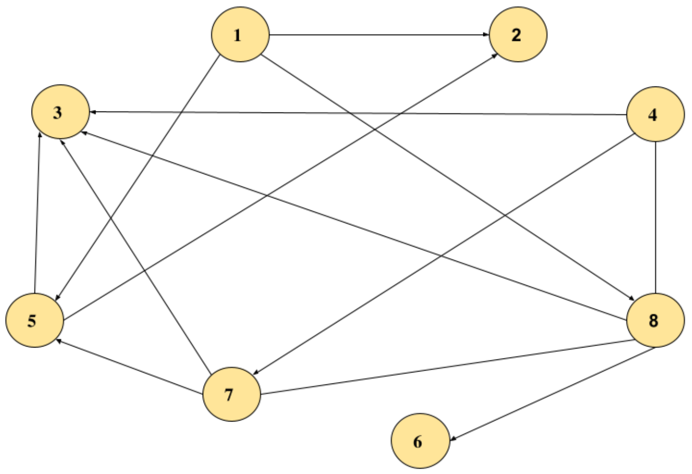


Table 3.3: Table 3

HITS		Eig	
Hub	Auth	Hub	Auth
4	3	8	3
7	8	4	8
8	7	1	7
1	5	7	4
5	2	2	6
2	4	3	5
3	6	5	2
6	1	6	1

Figure 3.11: Directed Graph for Participant Four

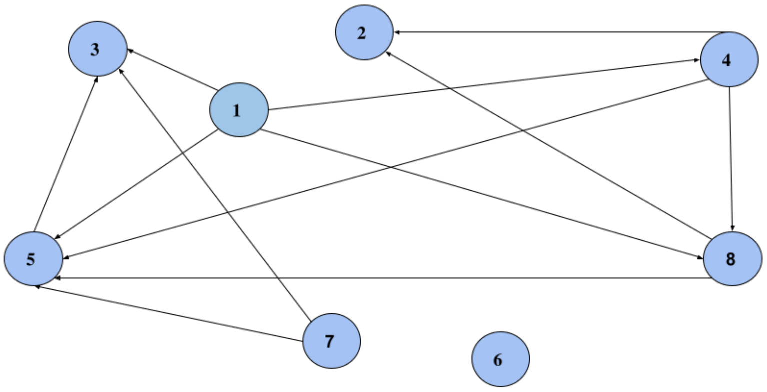


Table 3.4: Table 4

HITS		Eig	
Hub	Auth	Hub	Auth
1	5	1	3
4	3	2	1
7	8	3	2
8	2	4	4
5	4	5	5
2	1	6	6
3	6	7	7
6	7	8	8

Figure 3.12: Directed Graph for Participant Five

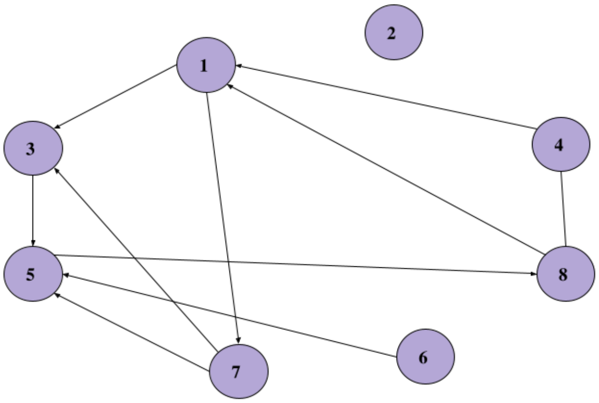


Table 3.5: Table 5

HITS		Eig	
Hub	Auth	Hub	Auth
7	5	8	1
1	3	4	8
3	7	1	5
6	2	7	3
2	4	5	4
4	8	3	7
5	1	6	2
8	6	2	6

Figure 3.13: Directed Graph for Participant Six

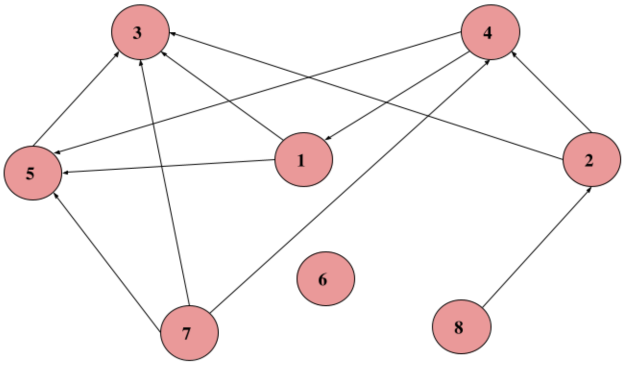


Table 3.6: Table 6

HITS		Eig	
Hub	Auth	Hub	Auth
7	3	2	3
2	5	8	5
1	4	1	1
5	2	3	2
4	1	4	4
8	6	5	6
3	7	6	7
6	8	7	8

Figure 3.14: Directed Graph for Participant Seven

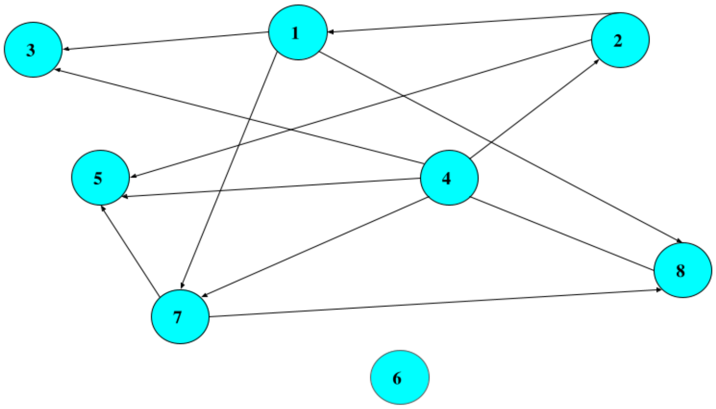


Table 3.7: Table 7

HITS		Eig	
Hub	Auth	Hub	Auth
4	5	4	5
1	8	1	8
7	7	8	4
2	3	2	3
3	2	7	7
5	1	3	2
6	4	5	1
8	6	6	6

Figure 3.15: Directed Graph for Participant Eight

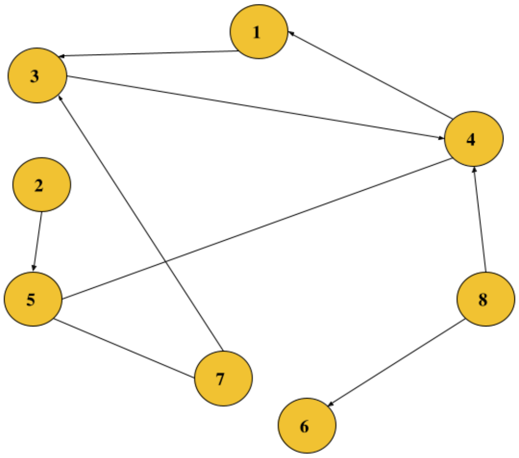


Table 3.8: Table 8

HITS		Eig	
Hub	Auth	Hub	Auth
7	5	5	4
4	3	7	5
2	1	4	3
1	2	2	1
5	4	3	7
8	6	8	2
3	7	1	6
6	8	6	8

Figure 3.16: Directed Graph for Participant Nine

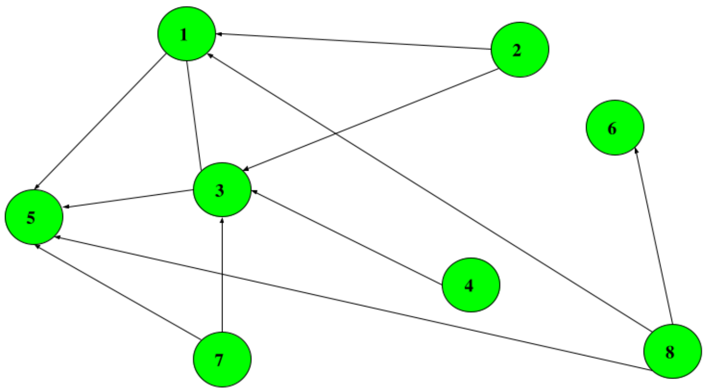


Table 3.9: Table 9

HITS		Eig	
Hub	Auth	Hub	Auth
8	5	8	5
1	3	1	1
7	1	2	3
3	4	3	2
2	8	4	4
4	2	5	6
5	6	6	7
6	7	7	8

Figure 3.17: Directed Graph for Group 1 - Female

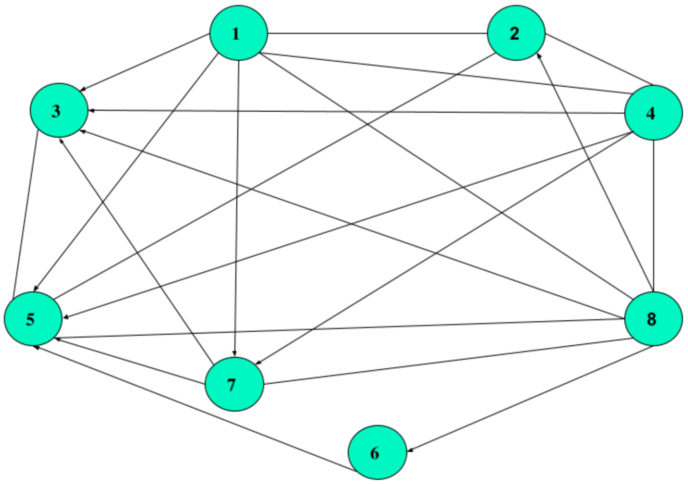


Table 3.10: Table 10

HITS		Eig	
Hub	Auth	Hub	Auth
8	5	8	5
1	3	4	3
4	8	1	2
7	2	2	8
2	7	7	4
5	1	5	1
3	4	3	7
6	6	6	6

Figure 3.18: Directed Graph for Group 2 - Male

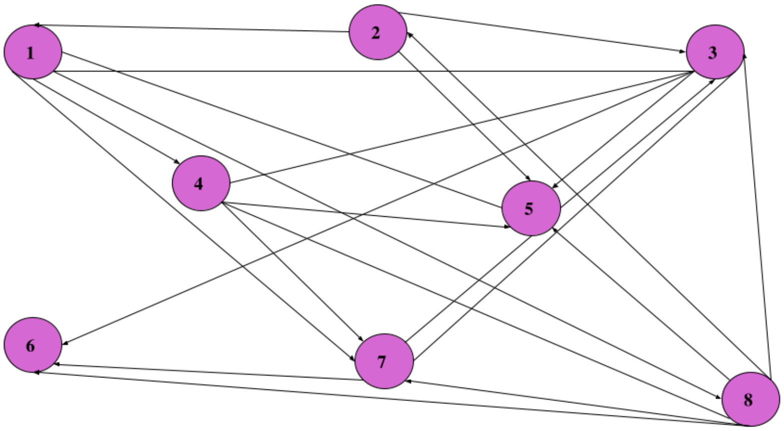


Table 3.11: Table 11

HITS		Eig	
Hub	Auth	Hub	Auth
8	5	1	5
1	3	8	3
3	7	3	6
4	6	4	7
7	4	2	1
2	1	5	4
5	8	7	8
6	2	6	2

The experimental results, for each participant and the groups, from PageRank are located in Appendix B.

3.4 Discussion of Our Findings

In this section we highlight some of the important observations noted from the experimental results of HITS, Dominant Eigenvector, and PageRank. Before diving deeper into these observations, we demonstrate how to analyze the results of the experiment by using Participant One as an example.

Figure 3.8 displays the digraph of Participant One, and Table 3.1 includes the rank of each node (social media category) on the graph as computed by HITS and Dominant Eigenvector. We see that HITS ranks node 7 (Beauty/Clothing/Self Care) as the most important hub and node 3 (Companies and Institutions) as the most important authority, while Dominant Eigenvector ranks node 2 (Climate and Animal Justice) as the most important hub and node 6 (Sports) as the most important authority. The results for Participant One from PageRank - found in Appendix B, Table B.1 - state that node 6 is the most important hub while node 3 is the most important authority. By the term “most important hub” we imply that this node was the most significant in terms of broadcasting information while "most important authority" refers to the node that receives the most information in the network. Now, we focus on the nodes deemed important by HITS. It intuitively makes sense that node 7 is the most important hub because it has 3 out-going edges: graphically, we see that node 7 has the maximum out-degree. Whereas node 3, the most important authority, has 3 in-coming edges: graphically, node 3 has the maximum in-degree. Another point to recall is that good hubs point to good authorities, and vice versa. Therefore, the results from HITS match our intuition regarding the digraph.

The fact that the methods do not agree on which nodes are the most important, in the case of Participant One, does not imply that our findings are incorrect. The mathematics associated with each method is significantly different, and therefore we expect differing results. In the case of some participants, the methods did agree on either the most important hub, authority or even both.

Since our intention was to provide participants with a recommendation based on our findings, we chose to recommend two nodes to each participant. These two nodes would be the most important hub, and the most important authority as determined by HITS. Participants could accept one of the two nodes (social media categories) or reject both. By rejection we imply that the participant chose one of the six other social media categories, as listed in section 3.2 because they deemed that category to be more significant to them. However, if the participant accepted any one of our recommendations, in some sense, our experiment succeeded. Thus, in the case of Participant One, they chose to accept our recommendation of node 3 which is Companies and Institutions. We went through the same process for each participant and offered our recommendation to them. Six out of the nine participants agreed with the results of our experiment and accepted that they would in fact be most interested in the social media category we predicted. Some of them were even surprised to see that the data, as well as conclusions, lined up so well in comparison to their real-world social media behavior.

With respect to the results from when the participants were grouped into the categories of female and male the conclusions were as follows. For Group 1 (female) all three methods agreed upon Education being the most important hub and Home/Culture/Lifestyle being the most important authority. For Group 2 (male) the methods determined the following:

HITS:

Most important hub = Education.

Most important authority = Home/Culture/Lifestyle.

Eigenvector:

Most important hub = Art.

Most important authority = Home/Culture/Lifestyle.

PageRank:

Most important hub = Education.

Most important authority = Companies and Institutions.

Chapter 4

Conclusions and Future Work

Ranking methods offer significant potential in developing and improving recommendation systems. Through a number of existing techniques, key insights into vertex prestige can assist in predicting user preferences on a number of applications. Such techniques are especially invaluable in the case of social media, retail, and e-commerce. We began with the goal of providing recommendations to participants, after having accessed their past consumption choices on Instagram, by designing an experiment. With the assistance of the three ranking methods, HITS, Dominant Eigenvector, and PageRank we were able to provide participants with suggestions related to what content they might be interested in on the app. We were partly successful in predicting what our participants prefer in terms of social media content with six out of nine participants accepting our recommendation.

There are a number of ways in which the experiment can be modified for further research. Since the experiment conducted in this project was confined to a particular environment, a natural direction for future work would be gathering and analysing a larger and more diverse data set. A different application can be considered, along with an intention to predict different commodities. Twitter, Netflix, Amazon, or any website can be a possible base to develop the model. Further, the data set can include several individuals from various locations if the experimental set-up allowed for a large data set.

Another possible direction for further exploration of this topic can be developing our own ranking algorithm. The experience acquired with this experiment can be a stepping stone to the creation of another method that computes vertex prestige. Analyzing non-symmetric Laplacians, as discussed in Chapter 2, and their properties is a clear starting point to building something new.

Bibliography

[1] Michele Benzi and Christine Klymko. On the limiting behavior of parameter-dependent network centrality measures. *SIAM Journal on Matrix Analysis and Applications*, 36(2):686–706, 2015.

[2] Phillip Bonacich. Factoring and weighting approaches to status scores and clique identification. *Journal of mathematical sociology*, 2(1):113–120, 1972.

[3] Fragkiskos D Malliaros and Michalis Vazirgiannis. Clustering and community detection in directed networks: A survey. *Physics reports*, 533(4):95–142, 2013.

[4] Mike Perkowitz and Oren Etzioni. Towards adaptive web sites: Conceptual framework and case study. *Artificial intelligence*, 118(1-2):245–275, 2000.

[5] Ioannis Pitas. *Graph-based social media analysis*, volume 39. CRC Press, 2016.

[6] Sebastiano Vigna. Spectral ranking. *Network Science*, 4(4):433–445, 2016.

Appendix A

Dominant Eigenvector and HITS: Detailed Numerical Results

Table A.1: Hub and Authority Ranking: Participant One

HITS				Eigenvector			
Node	Hub Score	Node	Authority Score	Node	Hub Score	Node	Authority Score
7	0.750133	3	0.607227	2	0.500895	6	0.456937
3	0.481641	5	0.544643	1	0.444612	5	0.456937
2	0.354689	6	0.544643	4	0.431606	3	0.443570
8	0.268493	1	0.194942	7	0.413861	8	0.425333
5	0.086196	2	0.000000	5	0.325859	1	0.334892
1	0.000000	4	0.000000	3	0.238825	4	0.245445
4	0.000000	7	0.000000	8	0.175036	7	0.179888
6	0.000000	8	0.000000	6	0.000000	2	0.000000

Table A.2: Hub and Authority Ranking: Participant Two

HITS				Eigenvector			
Node	Hub Score	Node	Authority Score	Node	Hub Score	Node	Authority Score
8	0.684689	5	0.668632	1	1.000000	3	1.000000
7	0.489996	3	0.494754	2	0.000000	1	0.000000
1	0.412833	7	0.462255	3	0.000000	2	0.000000
2	0.244086	2	0.249947	4	0.000000	4	0.000000
5	0.180611	6	0.178874	5	0.000000	5	0.000000
4	0.168747	1	0.000000	6	0.000000	6	0.000000
3	0.000000	4	0.000000	7	0.000000	7	0.000000
6	0.000000	8	0.000000	8	0.000000	8	0.000000

Table A.3: Hub and Authority Ranking: Participant Three

HITS				Eigenvector			
Node	Hub Score	Node	Authority Score	Node	Hub Score	Node	Authority Score
4	0.528201	3	0.644151	8	0.601501	3	0.740650
7	0.514517	8	0.499734	4	0.601501	8	0.370325
8	0.468595	7	0.351977	1	0.371748	7	0.370325
1	0.372527	5	0.313222	7	0.371748	4	0.228873
5	0.312920	2	0.242037	2	0.000000	6	0.228873
2	0.000000	4	0.165465	3	0.000000	5	0.228873
3	0.000000	6	0.165465	5	0.000000	2	0.141451
6	0.000000	1	0.000000	6	0.000000	1	0.000000

Table A.4: Hub and Authority Ranking: Participant Four

HITS				Eigenvector			
Node	Hub Score	Node	Authority Score	Node	Hub Score	Node	Authority Score
1	0.640832	5	0.695311	1	1.000000	3	1.000000
4	0.515409	3	0.434309	2	0.000000	1	0.000000
7	0.407254	8	0.416851	3	0.000000	2	0.000000
8	0.365125	2	0.317452	4	0.000000	4	0.000000
5	0.156578	4	0.231034	5	0.000000	5	0.000000
2	0.000000	1	0.000000	6	0.000000	6	0.000000
3	0.000000	6	0.000000	7	0.000000	7	0.000000
6	0.000000	7	0.000000	8	0.000000	8	0.000000

Table A.5: Hub and Authority Ranking: Participant Five

HITS				Eigenvector			
Node	Hub Score	Node	Authority Score	Node	Hub Score	Node	Authority Score
7	0.707107	5	0.788675	8	0.539345	1	0.458160
1	0.408248	3	0.577350	4	0.539345	8	0.458160
3	0.408248	7	0.211325	1	0.333333	5	0.458160
6	0.408248	2	0.000000	7	0.333333	3	0.458160
2	0.000000	4	0.000000	5	0.333333	4	0.283158
4	0.000000	8	0.000000	3	0.206011	7	0.283158
5	0.000000	1	0.000000	6	0.206011	2	0.000000
8	0.000000	6	0.000000	2	0.000000	6	0.000000

Table A.6: Hub and Authority Ranking: Participant Six

HITS				Eigenvector			
Node	Hub Score	Node	Authority Score	Node	Hub Score	Node	Authority Score
7	0.622747	3	0.711581	2	0.707107	3	0.834058
2	0.523681	5	0.493242	8	0.707107	5	0.417029
1	0.457467	4	0.435294	1	0.000000	1	0.208514
5	0.270185	2	0.232335	3	0.000000	2	0.208514
4	0.218831	1	0.083089	4	0.000000	4	0.208514
8	0.088217	6	0.000000	5	0.000000	6	0.000000
3	0.000000	7	0.000000	6	0.000000	7	0.000000
6	0.000000	8	0.000000	7	0.000000	8	0.000000

Table A.7: Hub and Authority Ranking: Participant Seven

HITS				Eigenvector			
Node	Hub Score	Node	Authority Score	Node	Hub Score	Node	Authority Score
4	0.698060	5	0.540224	4	0.658350	5	0.541127
1	0.625213	8	0.540224	1	0.462041	8	0.541127
7	0.349030	7	0.427472	8	0.429128	4	0.352719
2	0.000000	3	0.427472	2	0.301169	3	0.327594
3	0.000000	2	0.225502	7	0.279716	7	0.327594
5	0.000000	1	0.000000	3	0.000000	2	0.229911
6	0.000000	4	0.000000	5	0.000000	1	0.149861
8	0.000000	6	0.000000	6	0.000000	6	0.000000

Table A.8: Hub and Authority Ranking: Participant Eight

HITS				Eigenvector			
Node	Hub Score	Node	Authority Score	Node	Hub Score	Node	Authority Score
7	0.656539	5	0.844030	5	0.554011	4	0.568529
4	0.577350	3	0.449099	7	0.490083	5	0.539233
2	0.428525	1	0.293128	4	0.428250	3	0.403165
1	0.228013	2	0.000000	2	0.334223	1	0.342981
5	0.000000	4	0.000000	3	0.258354	7	0.325308
8	0.000000	6	0.000000	8	0.258354	2	0.000000
3	0.000000	7	0.000000	1	0.155860	6	0.000000
6	0.000000	8	0.000000	6	0.000000	8	0.000000

Table A.9: Hub and Authority Ranking: Participant Nine

HITS				Eigenvector			
Node	Hub Score	Node	Authority Score	Node	Hub Score	Node	Authority Score
8	0.561901	5	0.658627	8	1.000000	5	0.816497
1	0.420060	3	0.499658	1	0.000000	1	0.408248
7	0.420060	1	0.483221	2	0.000000	2	0.408248
3	0.414099	4	0.203777	3	0.000000	3	0.000000
2	0.356447	8	0.203777	4	0.000000	4	0.000000
4	0.181204	2	0.000000	5	0.000000	6	0.000000
5	0.000000	6	0.000000	6	0.000000	7	0.000000
6	0.000000	7	0.000000	7	0.000000	8	0.000000

Table A.10: Hub and Authority Ranking: Group 1

HITS				Eigenvector			
Node	Hub Score	Node	Authority Score	Node	Hub Score	Node	Authority Score
8	0.578307	5	0.487747	8	0.626154	5	0.568647
1	0.492399	3	0.432050	4	0.483725	3	0.463268
4	0.492399	8	0.395169	1	0.483725	2	0.396765
7	0.281678	2	0.371712	2	0.295931	8	0.304110
2	0.223225	7	0.334831	7	0.207446	4	0.272412
5	0.172173	1	0.277172	5	0.089868	1	0.272412
3	0.104480	4	0.277172	3	0.025153	7	0.237607
6	0.104480	6	0.123879	6	0.025153	6	0.085117

Table A.11: Hub and Authority Ranking: Group 2

HITS				Eigenvector			
Node	Hub Score	Node	Authority Score	Node	Hub Score	Node	Authority Score
8	0.504667	5	0.536117	1	0.501909	5	0.485375
1	0.443260	3	0.496563	8	0.465485	3	0.485375
3	0.413701	7	0.397958	3	0.396168	6	0.442455
4	0.371466	6	0.338624	4	0.387894	7	0.363710
7	0.314883	4	0.312661	2	0.341566	1	0.300095
2	0.286793	1	0.216291	5	0.263975	4	0.281088
5	0.241444	8	0.187080	7	0.194038	8	0.170829
6	0.000000	2	0.115883	6	0.000000	2	0.050212

Appendix B

PageRank: Detailed Numerical Results

Table B.1: PageRank Results: Participant One

PageRank			
Node	Hub Score	Node	Authority Score
6	0.179445	3	0.203264
1	0.158762	1	0.169139
4	0.158345	8	0.153561
7	0.157876	5	0.142433
3	0.119603	6	0.142433
2	0.112984	4	0.102374
5	0.074712	7	0.068991
8	0.038273	2	0.017804

Table B.2: PageRank Results: Participant Two

PageRank			
Node	Hub Score	Node	Authority Score
7	0.228704	3	0.321586
3	0.193802	5	0.207048
6	0.193802	7	0.145374
8	0.155258	6	0.101322
1	0.084401	2	0.066079
4	0.070370	8	0.052863
5	0.059631	1	0.052863
2	0.014031	4	0.052863

Table B.3: PageRank Results: Participant Three

PageRank			
Node	Hub Score	Node	Authority Score
8	0.247758	3	0.226588
4	0.146026	8	0.149312
1	0.139251	2	0.134905
7	0.121005	7	0.125737
3	0.094793	5	0.117878
6	0.094793	4	0.094303
2	0.094793	6	0.094303
5	0.061580	1	0.056974

Table B.4: PageRank Results: Participant Four

PageRank			
Node	Hub Score	Node	Authority Score
1	0.181624	3	0.298429
3	0.168958	5	0.188482
6	0.168958	2	0.141361
2	0.168958	8	0.104712
4	0.117354	4	0.078534
8	0.082583	1	0.062827
7	0.061504	7	0.062827
5	0.050062	6	0.062827

Table B.5: PageRank Results: Participant Five

PageRank			
Node	Hub Score	Node	Authority Score
2	1.000000	8	0.253968
1	0.000000	1	0.190476
3	0.000000	5	0.190476
4	0.000000	3	0.142857
5	0.000000	4	0.126984
6	0.000000	7	0.095238
7	0.000000	6	0.000000
8	0.000000	2	0.000000

Table B.6: PageRank Results: Participant Six

PageRank			
Node	Hub Score	Node	Authority Score
3	0.233063	3	0.332016
6	0.233063	5	0.169960
2	0.158103	2	0.142292
7	0.104709	4	0.110672
4	0.074243	1	0.102767
1	0.071712	8	0.047431
8	0.070268	7	0.047431
5	0.054838	6	0.047431

Table B.7: PageRank Results: Participant Seven

PageRank			
Node	Hub Score	Node	Authority Score
4	0.225528	4	0.205696
8	0.164021	5	0.166139
1	0.147508	8	0.166139
2	0.107278	1	0.120253
3	0.093155	3	0.110759
5	0.093155	7	0.110759
6	0.093155	2	0.080696
7	0.076200	6	0.039557

Table B.8: PageRank Results: Participant Eight

PageRank			
Node	Hub Score	Node	Authority Score
6	0.529412	4	0.315789
8	0.470588	3	0.210526
3	0.000000	5	0.210526
1	0.000000	1	0.157895
4	0.000000	7	0.105263
5	0.000000	2	0.000000
7	0.000000	6	0.000000
2	0.000000	8	0.000000

Table B.9: PageRank Results: Participant Nine

PageRank			
Node	Hub Score	Node	Authority Score
8	0.380330	5	0.304348
6	0.192989	3	0.246377
5	0.192989	1	0.202899
3	0.064016	8	0.057971
1	0.060472	4	0.057971
7	0.060472	2	0.043478
2	0.033669	6	0.043478
4	0.015063	7	0.043478

Table B.10: PageRank Results: Group 1

PageRank			
Node	Hub Score	Node	Authority Score
8	0.292202	5	0.329480
1	0.215048	3	0.210983
4	0.215048	2	0.199422
2	0.148837	4	0.086705
7	0.079617	1	0.086705
5	0.038304	8	0.046243
3	0.005472	7	0.034682
6	0.005472	6	0.005780

Table B.11: PageRank Results: Group 2

PageRank			
Node	Hub Score	Node	Authority Score
8	0.204090	3	0.194920
1	0.196967	5	0.175428
3	0.150403	6	0.172059
4	0.149217	1	0.130146
2	0.103681	7	0.123187
5	0.098590	4	0.098550
7	0.051657	8	0.072174
6	0.045395	2	0.033536