EXPLORING PATTERNS IN MATHEMATICS FACULTY HIRING WITH PERSISTENT HOMOLOGY

MAKENNA GREENWALT*, KELVIN LUU†, AND AHOORA TAMIZIFAR‡

Abstract. In the mathematics community, there is a strong link between the prestige of one's PhD program and one's likelihood of securing a faculty position. However, the web of faculty hiring pipelines is not well understood. In this work, we analyze the last 45 years of data from the Mathematics Genealogy Project to create a weighted network of PhD-granting institutions in the United States. For each pair of institutions, we calculate how "close" they are based on the number of graduates from one institution that are hired as faculty at the other. We then apply techniques in persistent homology to analyze the underlying patterns in the dataset. Finally, we extend this analysis by focusing on gender-based differences. Our results indicate an underlying hierarchy of institutions within mathematics faculty hiring networks, with connections differing significantly between men and women.

1. Introduction

For those who want a career in research, the preparation to become a faculty member begins long before the final year of graduate school. The list of publications, talks, and awards required to build a strong CV is amassed through an entire PhD program's worth of opportunities. A choice of thesis advisor can open or close doors in a subfield, depending on the strength of their professional connections. The institution one receives their PhD from (henceforth, their "PhD institution") should thus have influence on the institution one eventually becomes faculty at (henceforth, their "faculty institution"), if one becomes faculty at all.

Previous studies of faculty hiring networks have approached their analysis with network methods. For instance, a paper of Clauset et al. defines a prestige ranking on institutions with the property that, in most cases, hires move from higher ranked institutions to lower ranked ones [1]. Clauset et al. observe inequality in the hiring of men and women: if a woman and a man graduate from comparatively ranked institutions, the rank of the institution that hired the woman tends to be lower than the rank of the institution that hired the man.

A paper of FitzGerald et al. uses network centrality to determine a particular set of "elite" institutions that has collectively held a majority of authority centrality in their considered network over multiple time frames between 1950 to 2019 [2]. Here, high authority can be understood as the relative strength of an institution's ability to produce PhD students who move on to become faculty at prestigious institutions. FitzGerald et al. notice that women appear less likely to obtain a faculty position, though they note that the observed disparity is decreasing over time.

We extend this work by analyzing faculty hiring networks from the perspective of *persistent homology*, one of the more popular methods in the growing field of topological data analysis (TDA). Persistent homology tracks the birth and death of topological features (namely, homology classes) in a dataset to analyze its "shape." For our application to faculty hiring networks, persistent homology provides higher-dimensional information about gaps in connection between a potentially large set of institutions, rather than just between pairs. The *persistence* (i.e. survival time) of such gaps gives us a measure of their importance.

Our application of persistent homology to the network setting is heavily inspired by and informed by the work of Ignacio and Darcy in their analysis of migration and remittance networks using [3]. In particular, we augment the standard persistent homology workflow by encoding directional information into *directed clique complexes* [4, 5]. This allows us to investigate hierarchical patterns ("flow") in our hiring network.

In this report, we examine the faculty hiring connections within a fixed collection of 150 "top" institutions between 1980–2022. We inspect the "strongest" connections, as measured by the proportion of faculty

* *University of Oregon*

† *University of California, Los Angeles*

‡ *University of California, Irvine*

members at an institution who received their PhD from the other. We employ such measures of connection of strength to apply persistent homology to our network in 3 ways. In the first, we treat connections between institutions as a symmetric relationship, ignoring direction. In the second, we do consider direction. In the third, we restrict our attention solely to two-way, reciprocal connections, that is, those where institutions in the pair both hire from and produce hires for the other.

Our report is organized as follows. In section 2, we discuss properties of the dataset used in our project, and we present some preliminary analysis of our dataset. We then provide a brief overview of TDA in section 3, and detail the implementation(s) of persistent homology to our problem in section 4. We discuss our results in section 5. We conclude and discuss future direction in section 6.

2. Dataset Description and Exploration

In this study, we used data taken from "Temporal Dynamics of Faculty Hiring in Mathematics" by FitzGerald et al [2]. The data was scraped from the Mathematics Genealogy Project (`https://www.genealogy.math.ndsu.nodak.edu/index.php`), an online database of mathematical scientists who have been granted doctoral degrees. The website is designed to record "mathematical lineage" and contains, in particular, data on the institution one received their degree from, the year of their degree, their advisor, and any of their graduated advisees. When collecting the data, FitzGerald et al chose to consider only individuals that one, were granted their PhD during or after 1950 and two, received their PhD from one of the top 150 U.S. PhD-granting institutions, as identified by the combined 1998, 2010, and 2018 U.S. News Graduate School Rankings for Mathematics Programs. The data was collected in 2022.

For each individual meeting the criteria, the following was recorded: their name, the year they received their doctoral degree, the name of the degree institution, whether they became a faculty member in mathematics, the name of their faculty institution, the name of their advisor's PhD institution, and whether their advisor graduated from a "top school."

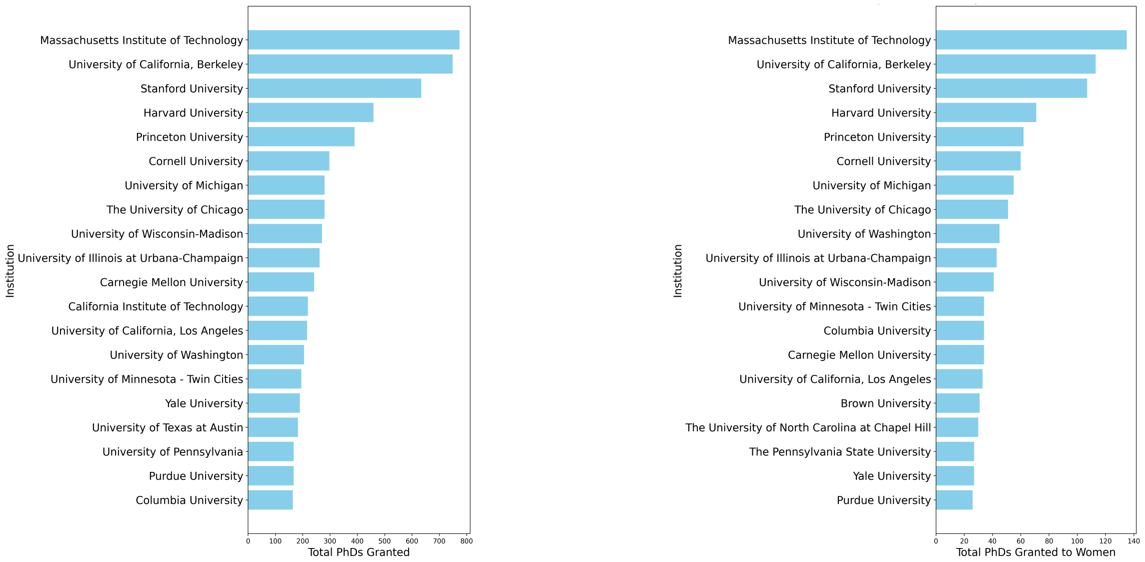
Since the Mathematics Genealogy Project does not provide information on current or past positions, whether or not an individual became a faculty member is inferred based on whether or not that individual is listed as having any PhD students. The institution from which that student received their PhD is then considered to be their advisor's faculty institution. If an individual had more than one PhD student, the institution at which they advised most students is considered their faculty institution, with the first chronological institution listed being considered in the event of a tie. Genders were assigned to names using the algorithm genderize.io. The names were then dropped from the dataset in order to anonymize the data.

Of interest to us in the final dataset are the year of degree, degree institution (i.e. PhD institution), faculty institution, and gender. As we are concerned with how institutions act as both producers *and* employers of mathematical scientists, we restrict ourselves to individuals whose faculty institution is also one of the 150 "top" institutions as determined by FitzGerald et al. This lets us focus on patterns in the transfer of individuals between these 150 schools.

We also restrict our consideration to individuals who graduated during or after 1980. This was a somewhat arbitrary cutoff meant to restrict our attention to modern hiring practice. As the faculty institution of an individual is determined from the degree institutions of their students, if an individual does not have any students, say, because they have only recently received their degree, then they will not have a listed faculty institution. In this case, that individual will not be present in our dataset. The 1980 cutoff was also decided to ensure we had enough data points in our dataset.

We examine the dataset discussed above in more detail. We first look at the total number of times an institution appears as a source institution, that is, the number of graduates produced by that institution *that move on to become faculty at one of our 150 institutions* between the years 1980 and 2022. Figure 1 shows the top 20 institutions in terms of such graduates. Notice that the data for women is similar to the overall dataset with some changes in ordering towards the bottom of the list.

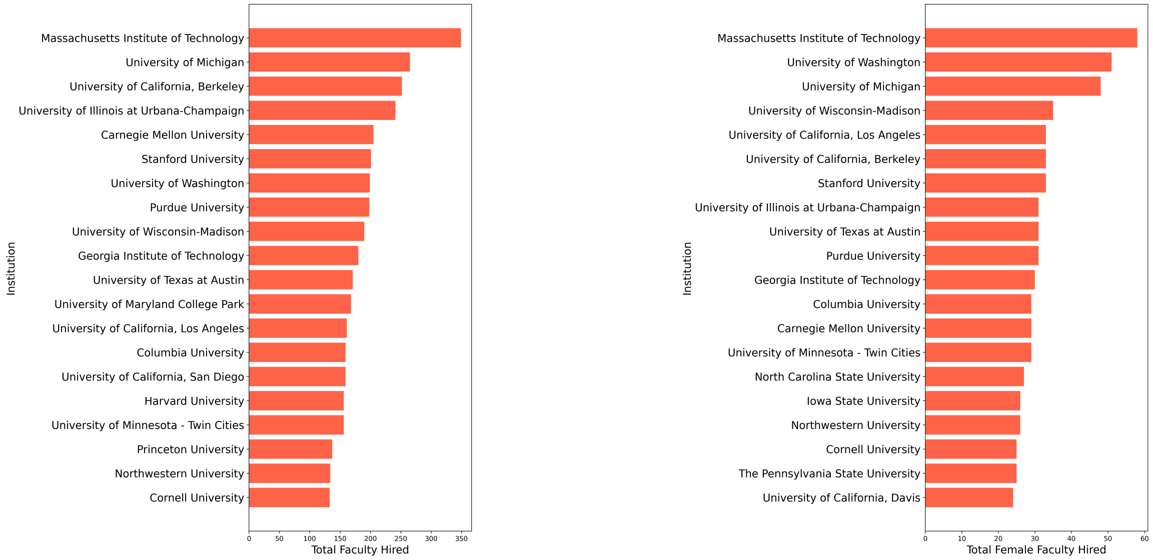
Now, we examine the total number of times an institution appears as a faculty destination, that is, the number of faculty members hired by that institution *who graduated from one of our considered institutions* between 1980 and 2022. Figure 2 shows the top 20 institutions in terms of faculty hired in our dataset. Here, we notice that several schools rank much higher on this list when only women are considered, including the University of Washington, the University of Wisconsin-Madison, and North Carolina State University;



(A) The top 20 institutions by total PhDs granted since 1980.

(B) The top 20 institutions by total PhDs granted to women since 1980.

Figure 1. Institutions that granted the most PhDs since 1980.



(A) The top 20 institutions by total faculty hired since 1980.

(B) The top 20 institutions by total female faculty hired since 1980.

Figure 2. Institutions that hired the most faculty since 1980.

similarly, Carnegie Mellon University and the University of Maryland, College Park rank much lower. Additionally, there are some institutions—including Harvard and Princeton—that no longer appear on the list at all.

3. Topological Data Analysis

We provide a brief introduction to topological data analysis (TDA) and its most widely used tool in applications, persistent homology. For a more detailed treatment, see [6, 7].

The theoretical foundations of TDA were developed in the early 2000s, but its methods have only gained widespread use in recent years due to advancements in computing. The core idea is to leverage topological and geometric features underlying the data to infer robust meaningful information about the structure. In this section, we introduce simplicial homology, define filtration and persistent homology, and conclude with a discussion of directed clique complexes and constructing simplicial complexes from networks.

3.1. Simplicial Homology.

The following definitions serve as a primer on simplicial homology. For a more comprehensive treatment, see [8, 9]. We will refer to simplicial homology as homology throughout the remainder of the report.

Definition 1. Given a set $\{x_0, x_1, \dots, x_n\} \subseteq \mathbb{R}^d$ of $n+1$ affinely independent points in Euclidean space, the n-dimensional simplex $\sigma = [x_0, x_1, \dots, x_n]$ is the convex hull spanned by these points, i.e.,

$$\sigma = \{t_0 x_0 + t_1 x_1 + \cdots + t_n x_n \in \mathbb{R}^d : t_0 + t_1 + \cdots + t_n = 1,\ t_0, t_1, \dots, t_n \geq 0\}.$$

The points that span the simplex are called *vertices*, and the simplices formed by subsets of these vertices are referred to as the *faces* of the simplex.

In low dimensions, a simplex is easy to visualize. A 0-simplex is a point. A 1-simplex, spanned by two points, is the line segment connecting them. A 2-simplex, spanned by three points, forms a triangle with these points as vertices. A 3-simplex, spanned by four points, forms a tetrahedron. Simplices are the building blocks of *simplicial complexes*, which are the structures we use to study the shape of data.

Definition 2. A simplicial complex K is a collection of simplices such that:

(1) Every face of a simplex of K is included in K.
(2) The intersection of any two simplices of K is either empty or a face of both simplices.

The dimension of a simplicial complex K is the maximum dimension of the simplices it contains.

Note that the union of the simplices of a simplicial complex K can be given the subspace topology inherited from the ambient space. Many applications of TDA deal with point cloud data (points in a metric space), and simplicial complexes are constructed from the points to extract topological features. Theoretical results, such as the nerve theorem [7], ensure that the topology of the simplicial complex reflects that of the point cloud. For our problem, we assign weights between schools to form a weighted network, from which we construct the simplicial complex. See Section 4 for the details of this construction.

We now turn our attention to computing homology groups for a given simplicial complex. Homology is an invariant from algebraic topology that detects "holes" in a topological space or simplicial complex. We restrict our homology computations to coefficients in the field $\mathbb{Z}_2$, as this is sufficient for our purposes.

Definition 3. Given a finite simplicial complex K and a non-negative integer k, the space of k-chains C_k on K with coefficients in $\mathbb{Z}_2$ is defined as the set of all formal sums of the k-simplices. More specifically, if $\{\sigma_1, \sigma_2, \dots, \sigma_p\}$ is the set of all k-simplices in K, then

$$C_k = \left\{ \sum_{i=1}^{p} a_i \sigma_i : a_i \in \mathbb{Z}_2 \right\}.$$

The space C_k with coefficients in $\mathbb{Z}_2$ forms a vector space with $\{\sigma_1, \sigma_2, \dots, \sigma_p\}$ as the basis.

Definition 4. Let K be a simplicial complex. The boundary $\partial_k(\sigma)$ of a k-simplex $\sigma = [v_0, v_1, \dots, v_k]$ of K is defined to be

$$\partial_k(\sigma) = \sum_{i=0}^{k} (-1)^i [v_0, \dots, \hat{v_i}, \dots, v_k],$$

where $[v_0, \dots, \hat{v_i}, \dots, v_k]$ is the $(k-1)$-simplex spanned by the set of all vertices of σ with v_i excluded.

Since the k-simplices form a basis for the space of all k-chains, the map ∂_k extends to a linear map from C_k to C_{k-1} called the boundary operator. It can be shown that the boundary operators satisfy the property

$$\partial_{k-1} \circ \partial_k = 0,$$

for all $k \geq 1$ [9]. This gives rise to the chain complex

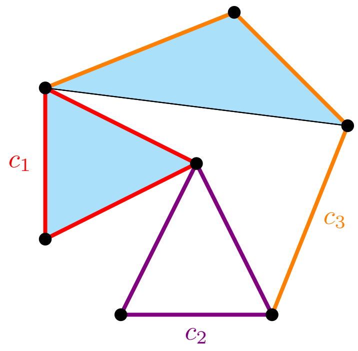


FIGURE 3. Examples of chains, cycles, and boundaries in a 2-dimensional simplicial complex: c_3 is a 1-chain, but not a 1-cycle or 1-boundary; c_2 is a 1-cycle, but not a 1-boundary; c_1 is a 1-cycle and 1-boundary.

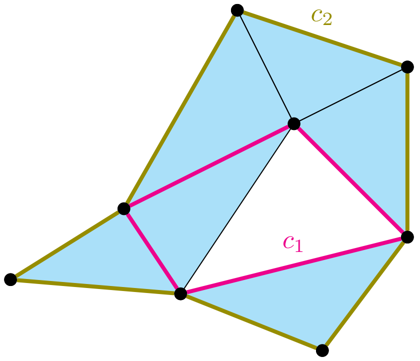


FIGURE 4. The 1-cycles c_1 and c_2 belong to the same homology class in H_1, as their difference is the boundary of the 2-chain represented by the union of triangles surrounded by c_1 and c_2.

$$\cdots \longrightarrow C_k \xrightarrow{\partial_k} C_{k-1} \xrightarrow{\partial_{k-1}} C_{k-2} \longrightarrow \cdots \longrightarrow C_1 \xrightarrow{\partial_1} C_0 \longrightarrow 0$$

of the simplicial complex K in $\mathbb{Z}_2$ coefficients. The kernel $\ker \partial_k = \{\sigma : \partial_k(\sigma) = 0\}$ of the boundary operator ∂_k is called the space of all *k-cycles* of K, and the image $\mathrm{im}(\partial_{k+1}) = \{\partial_{k+1}(c) : c \in C_{k+1}\}$ of the boundary operator ∂_{k+1} is called the space of all *k-boundaries* of K. Observe that any k-boundary is a k-cycle, i.e., $\mathrm{im}(\partial_{k+1}) \subseteq \ker \partial_k$. Figure 3 illustrates these notions.

Definition 5. Let K be a simplicial complex. The k-th homology group H_k of K is the quotient vector space

$$H_k = \frac{\ker \partial_k}{\mathrm{im}(\partial_{k+1})} .$$

The *k-th Betti number* β_k is defined as the rank of the k-th homology group, i.e., $\beta_k = \mathrm{rank}\,(H_k)$.

If the difference of two k-cycles is a k-boundary, then the cycles are said to be *homologous*. One can intuitively think of homologous k-cycles as surrounding the same k-dimensional hole. For an illustration of this, see Figure 4. The k-th Betti number represents the number of independent k-dimensional holes. Thus, the 0th Betti number counts the number of connected components, the 1st Betti number counts the number of independent loops, and the 2nd Betti number counts the number of independent cavities or voids.

3.2. **Persistent Homology.** With the foundations in place, we now introduce persistent homology.

Definition 6. A filtration of a simplicial complex K is a family of subcomplexes $\{K_i\}_{i\in I}$, where $I \subseteq \mathbb{R}$ and $K = \bigcup_{i\in I} K_i$, such that if $i \leq i'$, then $K_i \subseteq K_{i'}$.

For point cloud data, the most common filtrations are the Čech and Vietoris–Rips complexes [6]. Given a function $f : K \to \mathbb{R}$, such that $f(\tilde{\sigma}) \leq f(\sigma)$ whenever $\tilde{\sigma}$ is a face of the simplex σ, one can define a filtration by setting $K_r = f^{-1}((-\infty, r])$. This filtration is referred to as the *sublevel filtration*, and it is the one we use in our project.

As the filtration parameter increases, holes are formed and possibly filled in. Persistent homology records the filtration values at which homology classes are born and die. This multiset of birth-death pairs is represented as a persistence diagram. Figure 5 presents an example of this process.

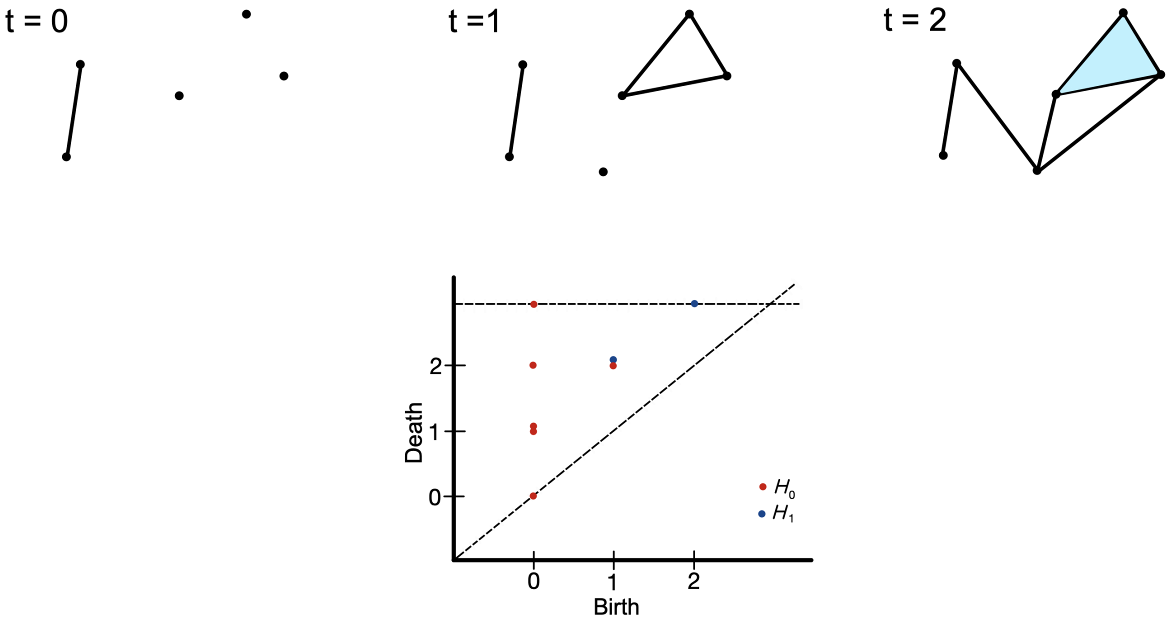


Figure 5. A filtration of a simplicial complex and its corresponding persistent diagram. Some points on the diagram are drawn on top of each other for ease of viewing. The simplicial complex for which persistent homology is calculated is the complex at filtration value 2. At filtration value 0, five connected components are born, along with an edge, which is represented by the red point on the diagonal dashed line. At filtration value 1, a new connected component and a 1-dimensional homology class are born, but both die at filtration value 2. The points on the horizontal dashed line represent homology classes that persist forever, and the number of such points determines the Betti number of the simplicial complex. There is one red point and one blue point on the horizontal dashed line, indicating that the complex is connected and has one loop.

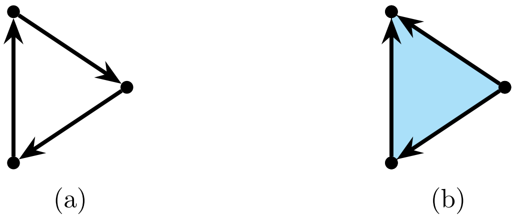
3.3. Simplicial Complexes for Undirected and Directed Networks.

Given an undirected network, one can construct a simplicial complex as follows: nodes and edges are included in the complex as 0-simplices and 1-simplices, respectively, and a k-simplex is included if edges between every pair of the $k+1$ nodes exist. Suppose instead that we have a directed network G. We would like to construct a simplicial complex that takes into account the directional information embedded in the network. We do this by considering *directed clique complexes*. For a more detailed exposition of clique complexes and their persistent homology computation, see [4, 5].

Definition 7. Let G be a directed network. A *directed k-clique* $\sigma = [v_0, \dots, v_k]$ is an ordered collection of $k+1$ vertices of G such that for each $i < j$, there is a directed edge (v_i, v_j) of G. We call v_0 the unique *source* of σ and v_k the unique *sink* of σ.

An ℓ-clique consisting of ℓ of the vertices of σ, where $\ell \leq k$, is called a *face* of σ if the vertices are ordered in the same manner.

To construct a directed clique complex from a given directed network, we regard the nodes and edges of the directed network as directed 0-cliques and 1-cliques, respectively. A k-clique is added if the collection of edges between the $k+1$ nodes have a unique source and sink. Figure 6 illustrates what constitutes a 2-clique and what does not. We convert a directed clique complex into a standard simplicial complex by removing the orientation of the edges. Persistent homology is then computed as usual.


(a)

(b)

Figure 6. An example and non-example of a directed 2-clique: there is no unique source and sink in (a), so it doesn't form a 2-clique, whereas (b) satisfies the clique conditions and is a 2-clique.

4. Model Implementation

We construct a weighted network with schools as nodes, assigning weights to the edges between them that capture the strength of their professional connections. A simplicial complex is created from this network as outlined in Section 3. For the filtration, we need to determine the filtration values at which simplices appear; the entries of the network's weighted adjacency matrix are used to calculate these values.

We would like the filtration to behave such that pairs of schools with the strongest connections are included first, while pairs with no connection are never included. Additionally, we wish to incorporate the directed nature of our data into the filtration. To best capture these intricacies, we employ three different methods for our analysis. Some of the code for this project was generated with ChatGPT [10].

4.1. **Undirected Network Method.**

Our first method is to define an undirected weighted network, from which we construct a simplicial complex. Inspired by [11], we take the naive approach of setting the weight wt_{AB} between two distinct schools A, B as

$$wt_{AB} = p(A,B) + p(B,A),$$

where $p(A,B)$ is the total number of faculty hired at B with PhDs from A, and similarly, $p(B,A)$ is the total number of faculty hired at A with PhDs from B.

The main issue with this approach is that it does not account for the size of the schools. Larger schools grant more degrees and hire more faculty than smaller schools; therefore, pairs of large schools will likely have higher weight values compared to pairs of small schools. This does not adequately reflect hiring practices and biases within institutions, so we would like a weight that considers school size as a factor.

In our second approach, we account for school size by taking proportions and defining the weight wt_{AB} between two distinct schools A, B as

$$wt_{AB} = \frac{p(A,B)}{p(A)} + \frac{p(B,A)}{p(B)},$$

where $p(A)$, $p(B) > 0$ are the total number of graduated PhD students from A and B that have found faculty positions, respectively. In other words, the weight between A and B is the sum of the proportion of PhD graduates from A, who became faculty, that have been hired at B and the proportion of PhD graduates from B, who became faculty, that have been hired at A.

Although this resolves the issue of mainly pairs of large schools having strong connections, some institutions with only a few PhD graduates who have secured faculty positions may exhibit a very strong connection to the schools where these graduates have been hired.

For example, we observe in our dataset that the University of Arkansas has only had one PhD graduate secure a faculty position since 1980, and that individual was hired at MIT. As a result, this connection will appear very strong in our analysis.[1]

To address the issue in this approach, we instead calculate the proportions by dividing by the number of faculty hired, setting the weight wt_{AB} as

$$wt_{AB} = \frac{p(A,B)}{n(B)} + \frac{p(B,A)}{n(A)} \,,$$

where $n(A)$, $n(B) > 0$ are the total number of faculty hired at A and B, respectively. Thus, the weight between A and B is the sum of the percentage of faculty hired at B with PhDs from A and the percentage of faculty hired at A with PhDs from B. Since schools tend to hire faculty frequently—and in fact, all the schools in our dataset have hired more than one faculty member since 1980—this weight assignment does not face the same issue as approach 2. This is the weight function used to construct the undirected network of schools from which we build our simplicial complex for the rest of this report.

Since we would like stronger connections to appear first in the filtration, we transform the weighted adjacency matrix of this network as follows. Let M be the maximum weight value. We define a weight matrix W and set $W_{AB} = M + 0.01 - wt_{AB}$. The diagonal entries of W are then set to zero, and any off-diagonal entry with a value of $M + 0.01$ is set to infinity. This ensures that all the schools appear at the beginning of the filtration, and edges between schools A, B such that $wt_{AB} = 0$ never appear in the filtration.

Thus, schools are represented as 0-simplices and appear at filtration value 0, while edges between schools are 1-simplices and appear at filtration value W_{AB} for schools A and B. The 2-simplex formed by three schools and their edges appears once all edges are connected, i.e., the maximum of the filtration values of their edges. The persistence diagrams generated from this method are presented in Section 5.

4.2. **Directed Clique Complex Method.** Our second method is to define a directed network from which we construct a directed clique complex. For two distinct institutions A and B, we define the weight between A and B as

$$wt_{AB} = \frac{p(A,B)}{n(B)} \,,$$

whenever $n(B) > 0$. A directed clique complex is formed from this directed network as outlined in 3.3. Then, we create a weight matrix W as in Section 4.1. Note that this matrix is not symmetric. As for the filtration, the 0-simplices are included like before; a 1-simplex between schools A and B is included at filtration value $\min\{W_{AB}, W_{BA}\}$; a 2-simplex with three schools as its vertices is included at the filtration value the last 1-simplex that forms the 2-clique appears. This method and the persistence diagrams corresponding to it are generated using the package Flagser [4].

4.3. **The Six-Edge Method.** Our third method also incorporates direction, but emphasizes reciprocated connections between schools. The same weight matrix from Section 4.2 is used for the filtration, and the 0-simplices are added like before. However, a 1-simplex between schools A and B is only added at filtration value $\min\{W_{AB}, W_{BA}\}$ if both W_{AB} and W_{BA} are finite, i.e., each school has hired at least one faculty member who earned their PhD at the other school. A 2-simplex with schools A, B, C is added at filtration value $\max\{W_{AB}, W_{AC}, W_{BA}, W_{BC}, W_{CA}, W_{CB}\}$ if all the six values are finite. In other words, a triangle formed by three schools is filled in if all directed six edges (two directions for each of the 3 sides) are present in the network. There are far fewer connections in this method, and as a result, more infinite connected components and 1-dimensional homology classes appear on the persistence diagrams.

5. Results

We start our analysis by applying the undirected network method to our data.

Table 1 shows the strongest 20 connections identified by our undirected network algorithm. Recall that lower filtration values indicate stronger connections in faculty hiring. In terms of weight value, the connection that stands out is between the University of Maryland, College Park, and Howard University. This connection was identified as ”strong” by the algorithm because only nine individuals in our dataset were hired at Howard

[1] No MIT graduates have become faculty at the University of Arkansas since 1980, so this connection is one-sided.

School Pair	Filtration Value
Howard University; University of Maryland College Park	0.0100
Stanford University; University of California, Berkeley	0.1325
Stanford University; Massachusetts Institute of Technology	0.2011
University of California, Berkeley; Massachusetts Institute of Technology	0.2034
Harvard University; Massachusetts Institute of Technology	0.2103
Rutgers University, Newark; Columbia University	0.2322
Stanford University; Princeton University	0.2438
University of Massachusetts Amherst; Massachusetts Institute of Technology	0.2457
Princeton University; Massachusetts Institute of Technology	0.2626
Clemson University; Virginia Polytechnic Institute and State University	0.2667
Stanford University; Cornell University	0.2668
University of California, Berkeley; University of Hawaii	0.2669
University of California, Berkeley; University of California, Santa Cruz	0.2687
Boston University; Massachusetts Institute of Technology	0.2687
Stanford University; California Institute of Technology	0.2696
Stanford University; Harvard University	0.2723
Cornell University; Wesleyan University	0.2726
Southern Methodist University; Rice University	0.2726
Wesleyan University; Massachusetts Institute of Technology	0.2726

Table 1. The top 20 pairs of schools with the strongest connections and the corresponding filtration values at which they appear in the simplicial complex.

University between 1980 and 2017, four of which obtained their PhDs at the University of Maryland. No PhD recipients from Howard University were hired at the University of Maryland.

Overall, the rest of the list from Table 1 aligns with our understanding of strong mathematical connections: top STEM schools are prominent, often alongside other prestigious institutions. Some less esteemed schools appear early in the dataset for reasons similar to Howard; for example, only six faculty members were hired at the University of Vermont during since 1980, so the University of Vermont has stronger connections with a few large schools than one might expect. It is also worth noting that several unexpected strong connections are close to one another geographically, such as Maryland - Howard, Rutgers, Newark - Columbia, UC Berkeley - UC Santa Cruz, and Rice - Southern Methodist. The remaining connections are illustrated in Figure 7, where 1-simplices are added as the filtration parameter increases.

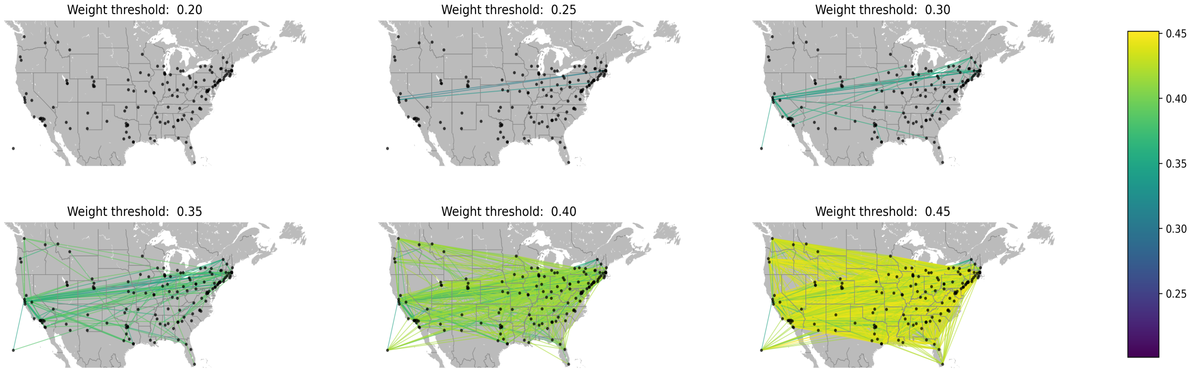


Figure 7. Filtration map of the undirected network method. Stronger connections appear first in the filtration. The dot in the lower-left corner represents the University of Hawaii.

Figure 8 presents the persistence diagram generated using the undirected network method. There is one connected component in H_0 and two classes in H_1 at infinity, but no other prominent features. We will use this diagram as a "base case" for our later analyses.

We now extend our analysis to address gender-related issues by modifying the weight calculation as follows: for two institutions A and B, we define

$$wt_{AB} = \frac{w(A,B)}{n(B)} + \frac{w(B,A)}{n(A)},$$

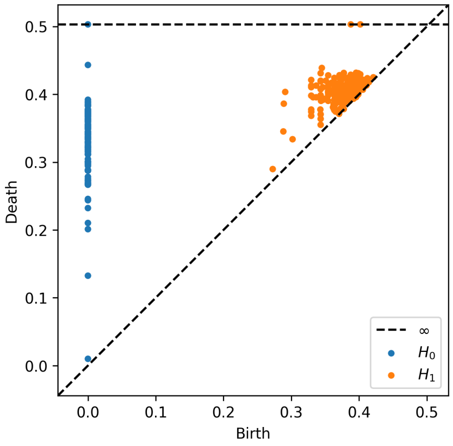


Figure 8. Persistence diagram of the undirected network method. Betti numbers: [1, 2].

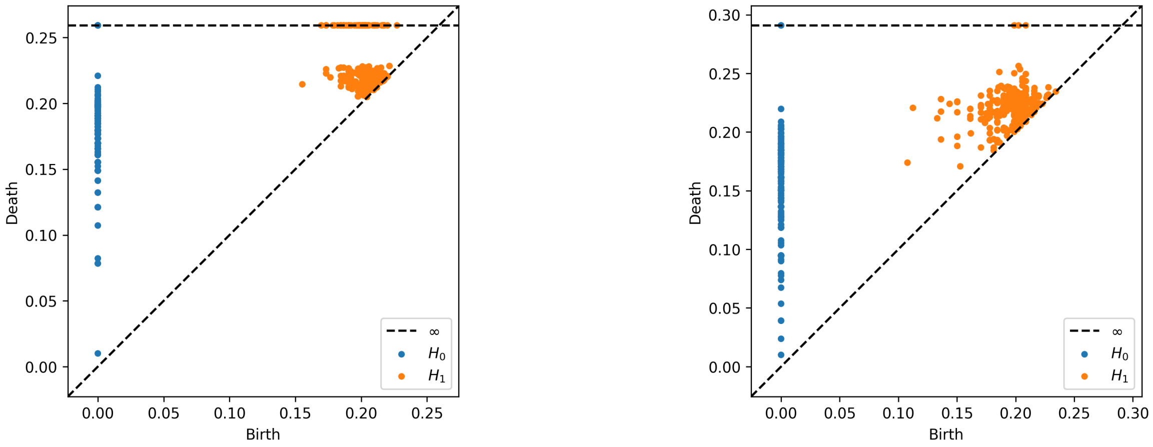
where $w(A, B)$ is the total number of female faculty hired at B with PhDs from A, and similarly, $w(B, A)$ is the total number of female faculty hired at A with PhDs from B. Note that $n(A)$ and $n(B)$ are the total number of faculty hired at institutions A and B, respectively, and *not* the total number of women hired.

This choice of divisor ensures that overall department size is reflected in our scaling. In this calculation, male-dominated institutions cannot appear early in the filtration, so our top connections will be between schools that regularly graduate and hire women. Furthermore, dividing by the number of female faculty hired presents the following issue: if an institution has hired only a few women, then it will have a strong connection with the select schools it has hired from, which may not accurately represent the strength of a hiring pipeline. The strongest connections for both women and men are presented in Table 2.

Strongest Connections for Women	**Strongest Connections for Men**
Howard University; University of Maryland College Park	University of California, Berkeley; Stanford University
University of Montana; Oregon State University	Massachusetts Institute of Technology; Stanford University
Claremont Graduate University; University of California, Davis	Rutgers University, Newark; Columbia University
Purdue University; New Mexico State University	Howard University; University of Maryland College Park
University of California, Merced; University of Maryland College Park	Massachusetts Institute of Technology; University of California, Berkeley
University of Connecticut; Howard University	Massachusetts Institute of Technology; Harvard University
George Mason University; Rutgers University, Newark	Massachusetts Institute of Technology; University of Massachusetts Amherst
University of Washington; Rutgers University, Newark	University of California, Berkeley; University of Hawaii
Howard University; University of Virginia	Stanford University; Princeton University
University of Alabama-Birmingham; University of Michigan	Cornell University; Wesleyan University
University of Alabama-Birmingham; University of California, Davis	Massachusetts Institute of Technology; Boston University
Massachusetts Institute of Technology; Wesleyan University	University of California, Berkeley; Cornell University
Rice University; Wesleyan University	Stanford University; California Institute of Technology
University of Pittsburgh; University of Nevada, Las Vegas	Massachusetts Institute of Technology; University of Vermont
Tulane University; The University of Memphis	University of California, Los Angeles; University of Nevada, Las Vegas
Massachusetts Institute of Technology; Brandeis University	University of California, Los Angeles; University of Vermont
University of California, Berkeley; The University of Memphis	The Ohio State University; University of Vermont
Dartmouth College; University of California, San Diego	Brandeis University; The University of Chicago
Tufts University; Massachusetts Institute of Technology	Brandeis University; Yale University
University of Texas at Austin; University of Montana	University of California, Santa Barbara; University of Nevada, Las Vegas

Table 2. The top 20 pairs of schools with the strongest connections for women and men.

Restricting by gender produces a noticeable difference in the output of our algorithm. While many of the strongest connections in the overall data were between "top" schools, no such pairings are found when the analysis is restricted to women. However, these pairings become more prominent when the data is restricted to men. This suggests a significant difference in hiring dynamics based on gender. There also appears to be no direct correlation between school pairings at the top of the women's list and their geographical proximity.



(A) Restricted to women. Betti numbers: [5, 67].

(B) Restricted to men. Betti numbers: [2, 3].

Figure 9. Persistence diagrams of the undirected method applied to gender specific data

Figure 9 displays the persistence diagram generated using the undirected network method on data restricted to women and men. Unsurprisingly, the two diagrams are quite different from each other. The points representing H_1 are more scattered in the men's dataset and clustered together in the women's.

The Betti numbers provide a more quantitative measure of the gender differences captured by the diagrams. The simplicial complex constructed from data restricted to men has two connected components and three independent H_1 classes, while the simplicial complex constructed from data restricted to women has five connected components and 67 independent H_1 classes. This is expected, as the undirected network with data restricted to women has fewer edges than that for men, resulting in fewer holes being filled.

The two connected components when restricting to men are 1) San Diego State University and 2) all other universities. The five connected components in the women's dataset are 1) University of Miami, 2) University of Louisville, 3) Ohio University, 4) University of Vermont, and 5) all other universities. This indicates that no women who received PhDs from these four institutions became faculty, and no women faculty were hired at those institutions. This is understandable for the University of Vermont (small mathematics program) and the University of Louisville (relatively new PhD program), but unusual for Ohio University and the University of Miami.

We now incorporate direction into our analysis by using the methods outlined in Sections 4.2 and 4.3. The strongest connections overall and for women are presented in Table 3.

Most of the strong connections identified in the undirected approach are represented here as one-sided connections, both for women and overall. Unsurprisingly, each arrow on this list indicates a decrease in mathematical prestige, reflecting the hierarchical nature of mathematics academia explored in [2].

Recall that in the six-edge method, two schools are connected if and only if there is a two-sided connection between them. The filtration values of these connections are represented in Figure 10.

The persistence diagrams for both the six-edge method and the directed clique complex method are displayed in 11. There are significantly more infinite H_0 and H_1 classes with the six-edge method, indicating that many of the connections from the undirected network method are one-sided.

The persistence diagram for the directed clique complex method looks very similar to the one generated by the undirected network method. They both have the same number of classes at infinity with similar birth times.[2] Furthermore, the patterns of points with finite death time in the two persistence diagrams are notably similar. This suggests that there are few 2-simplices in the undirected network method that do not form 2-cliques. In the context of mathematics faculty hiring, the presence of a clique indicates an

[2] The diagram for the directed clique complex method is produced using Flagser, which does not graph a single connected component at infinity.

Strongest Directed Connections for All	**Strongest Directed Connections for Women**
University of Maryland College Park → Howard University	University of Maryland College Park → Howard University
Columbia University → Rutgers University, Newark	Oregon State University → University of Montana
Massachusetts Institute of Technology → University of Massachusetts Amherst	University of California, Davis → Claremont Graduate University
University of California, Berkeley → University of Hawaii	Purdue University → New Mexico State University
Massachusetts Institute of Technology → Boston University	University of Maryland College Park → University of California, Merced
Massachusetts Institute of Technology → Wesleyan University	University of Connecticut → Howard University
Rice University → Southern Methodist University	George Mason University → Rutgers University, Newark
Cornell University → Wesleyan University	University of Washington → Rutgers University, Newark
University of California, Berkeley → University of California, Santa Cruz	University of Virginia → Howard University
Virginia Polytechnic Institute and State University → Clemson University	University of California, Davis → University of Alabama-Birmingham
University of California, Berkeley → University of California, Davis	University of Michigan → University of Alabama-Birmingham
The University of Chicago → Brandeis University	Massachusetts Institute of Technology → Wesleyan University
Massachusetts Institute of Technology → Brandeis University	Rice University → Wesleyan University
The Ohio State University → University of Vermont	Tulane University → The University of Memphis
University of Maryland College Park → University of Vermont	University of California, Berkeley → The University of Memphis
University of California, Santa Barbara → University of Nevada, Las Vegas	Massachusetts Institute of Technology → Brandeis University
University of California, Los Angeles → University of Nevada, Las Vegas	University of Pittsburgh → University of Nevada, Las Vegas
Massachusetts Institute of Technology → Northeastern University	University of California, San Diego → Dartmouth College
University of California, San Diego → University of Vermont	Massachusetts Institute of Technology → Tufts University
University of California, Berkeley → The University of Memphis	Cornell University → University of Montana

Table 3. The top 20 pairs of schools with the strongest directed connections overall and for women.

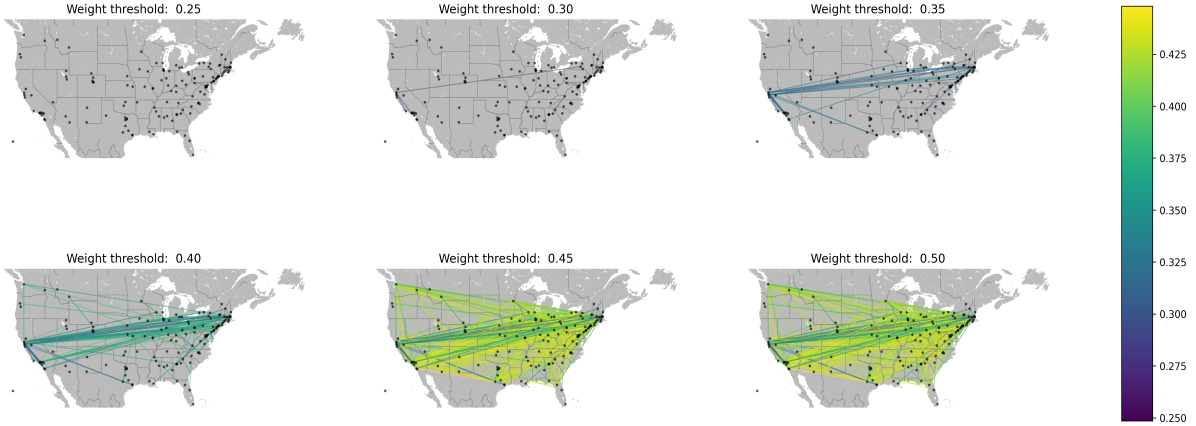
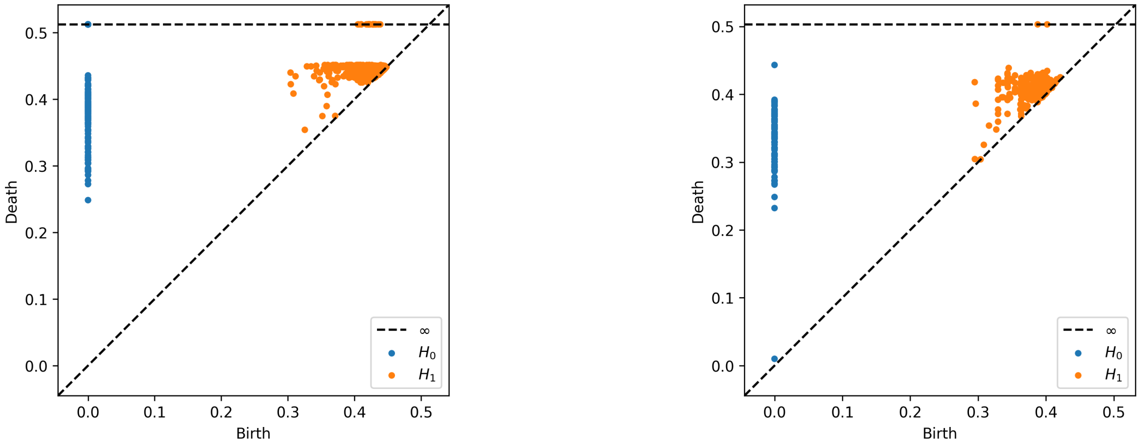


Figure 10. Filtration map of the six-edge method. Stronger connections appear first in the filtration. The dot in the lower-left corner represents the University of Hawaii.

implied hierarchy among three institutions. The similarity between the two diagrams signifies that the vast majority of these connections are hierarchical, reinforcing the picture painted in [2] of mathematics as a prestige-hierarchical field.

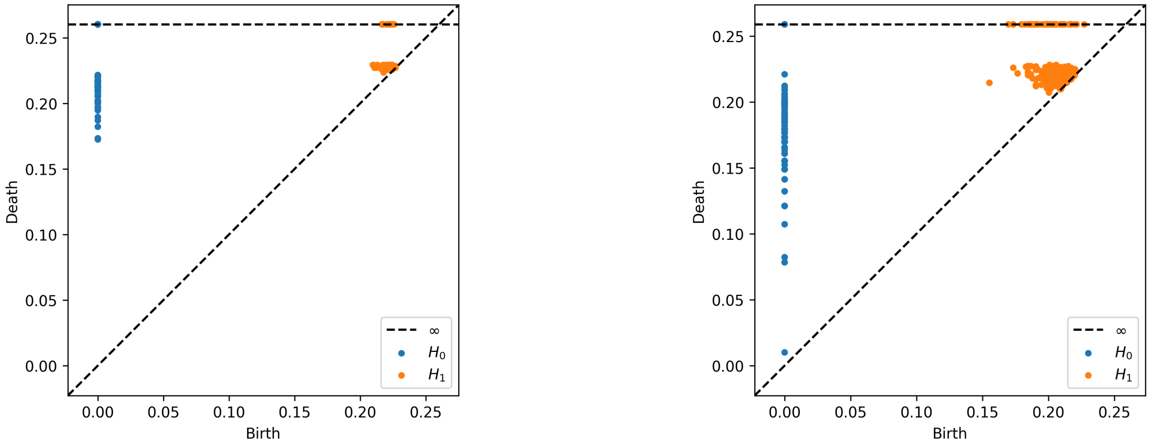
Figure 12 displays the persistence diagrams generated using the six-edge and directed clique complexes methods on data restricted to women. The diagram for the six-edge method is closely clustered, featuring 95 connected components at infinity, indicating that there are few reciprocated connections when considering only women. In contrast, the diagram for the directed clique complex method closely resembles that of the undirected network method. Together, these diagrams suggest that the few connections that do appear are often one-sided, with the 2-simplices having a unique source and sink, reflecting a hierarchical nature. We hypothesize that this reflects a prestige-based hierarchy analogous to that found in the overall data.



(A) Persistence diagram of the six-edge method. Betti numbers: [5, 73].

(B) Persistence diagram of the directed clique complex method. Betti numbers: [1, 2].

FIGURE 11. Persistence diagrams of the methods that incorporate direction



(A) Persistence diagram of the six-edge method for women. Betti numbers: [95, 22].

(B) Persistence diagram of directed clique complex method for women. Betti numbers: [5, 73].

FIGURE 12. Persistence diagrams of the methods that incorporate direction for women

6. Conclusion

The interconnected web of faculty hiring pipelines in mathematics is both anecdotally and statistically known to rely heavily on institutional prestige. Our work analyzes faculty hiring data in mathematics by creating a weighted network of connections between schools and applying persistent homology. We implemented several techniques to account for direction in our network and applied these algorithms to both gendered and ungendered versions of our data. Our resulting persistence diagrams supported the conclusion that there is a hierarchy of institutions that play a large role in hiring and suggested that many of the major hiring pipelines among top schools are not as accessible to women.

7. Limitations and Future Work

Our study has several limitations that stem from both the data we inherited and the nature of the Mathematics Genealogy Project. For example, it is possible to hold a faculty position without advising any students, particularly in short-term positions, but these individuals would not show up in our dataset. As a result, institutions that have more short-term faculty or low faculty retention rates may be misrepresented in the data. Furthermore, since it is normal in mathematics for one to hold a postdoctoral position before becoming faculty, there is often a large gap between the PhD year listed in the data and the graduation of one's first advisee. This means that the last ten years of data are likely incomplete compared to the trends established in the 1980s and 1990s. This can have a notable effect on programs created in the last twenty years, such as the University of Louisville and the University of Nevada, Reno.

We also found that the choices made by FitzGerald et al. in the construction of their dataset were not ideal for the analysis we wanted to perform. In particular, because they choose the institution an individual works at by selecting, in some sense, the one where they have worked the *longest*, we do not think the data was ideal for understanding the impact of their PhD institution on their hiring. There may have been some impacts from institutions they worked at between their obtaining their PhD and taking on their position at the listed institution. As we use the *graduation year* of an individual to determine which time-frame they belong to, not their *hiring year*, we did not believe this choice would accurately represent hiring practices during any given time-frame either. To help remedy this, we are in the process of constructing our own dataset by querying the Mathematics Genealogy API. Once complete, we should be able to perform a temporal analysis by restricting our new dataset to 15- or 20-year periods.

The main difference in how we are constructing our dataset is that we choose the *first* institution that an individual becomes faculty following their PhD. We believe this makes the most sense as a representation of an individual's hiring outcomes in a way that minimizes the influence of post-PhD placements. We select this by choosing the institution of their earliest graduated student. If there are conflicts—for example, if an individual had two students who graduated in the same year from different institutions—we resolve them manually as follows. In all cases we've encountered, conflicts have been between two institutions. If both of the conflicting institutions were really institutions (so, not blank entries), we selected the one that appeared the most often as the institution that an individual's advisees received their degree from. If there was a tie here, we would look up the individual online to see there was information, like from a CV, that could resolve it. If none of the above worked, we created two rows for the individual, one for each institution in conflict. Any advisees with blank institution or year were not considered. If an individual only had such advisees, they were thrown out as well.

Due to a lack of resources, we have been unable to genderize our dataset thus far. We hope to find an effective algorithm in the near future so that we can combine our temporal and gender-based analyses.

We also hope to examine homology classes with high persistence by computing a *cycle representative* of the class at various times to understand their composition.

Since Mathematics Genealogy Project includes a classification of mathematical subfields, a future direction could be to see if persistent cycles or clusters have any relation to those subfields. Another direction would be to add undergraduate-graduate transition data and then use a multilayered network to analyze long-term pipelines. However, this data is not readily available, and persistent homology techniques for multilayered networks are still relatively new.

Acknowledgments

This project was funded by National Science Foundation Research and Training Grant DMS-2136090. We thank our wonderful mentors, Mason Porter, Sarah Tymochko, and Sidhanth Raman for their unwavering support during our research. Thank you to Dr. Mitchel T. Keller and the Mathematics Genealogy Project for giving us access to their API. Finally, we express our sincere gratitude to those who have implemented TDA algorithms in publicly accessible code.

References

[1] Aaron Clauset, Samuel Arbesman, and Daniel B. Larremore. Systematic inequality and hierarchy in faculty hiring networks. *Science Advances*, 1(1):e1400005, 2015.

[2] Cody FitzGerald, Yitong Huang, Katelyn Plaisier Leisman, and Chad M. Topaz. Temporal dynamics of faculty hiring in mathematics. *Humanities and Social Sciences Communications*, 10(1):247, May 2023.

[3] Paul Samuel P. Ignacio and Isabel K. Darcy. Tracing patterns and shapes in remittance and migration networks via persistent homology. *EPJ Data Science*, 8(1):1, December 2019.

[4] Daniel Lütgehetmann, Dejan Govc, Jason P. Smith, and Ran Levi. Computing persistent homology of directed flag complexes. *Algorithms*, 13(1), 2020.

[5] Michael W. Reimann, Max Nolte, Martina Scolamiero, Katharine Turner, Rodrigo Perin, Giuseppe Chindemi, Paweł Dłotko, Ran Levi, Kathryn Hess, and Henry Markram. Cliques of neurons bound into cavities provide a missing link between structure and function. *Frontiers in Computational Neuroscience*, 11(48), June 2017.

[6] Frédéric Chazal and Bertrand Michel. An introduction to topological data analysis: Fundamental and practical aspects for data scientists. *Frontiers in Artificial Intelligence*, 4, 2021.

[7] Tamal Krishna Dey and Yusu Wang. *Computational Topology for Data Analysis*. Cambridge University Press, 2022.

[8] Allen Hatcher. *Algebraic topology*. Cambridge University Press, Cambridge, 2002.

[9] James R. Munkres. *Elements Of Algebraic Topology*. CRC Press, 2018.

[10] OpenAI. Chatgpt: Language model, 2023.

[11] Abigail Hickok, Benjamin Jarman, Michael Johnson, Jiajie Luo, and Mason A. Porter. Persistent homology for resource coverage: A case study of access to polling sites. *SIAM Review*, 66(3):481–500, 2024.