Discovering research archetypes and its variability with gender and grant income among computer scientists

Kanika Narang[1], Hari Sundaram[1], Snigdha Chaturvedi[2], Austin Chung[1],

1 Department of Computer Science, University of Illinois, Urbana-Champaign, Urbana, Illinois, United States
2 Department of Computer Science, University of California, Santa Cruz, Santa Cruz, California, United States

*knarang2@illinois.edu

Abstract

In this paper, we aim to discover archetypical patterns of research interests evolution among Academics. In our work, an archetype comprises of *progressive stages* of distinct research *behavior*. We introduce a novel Gaussian Hidden Markov Model (G-HMM) cluster model to identify archetypes of evolutionary patterns. G-HMMs allow for: behavioral variation and different evolutionary rates; impose constraints on how individuals can evolve; are interpretable.

Our model identifies four distinct archetypes for Computer Science researchers: *Steady*, *Diverse*, *Evolving* and *Diffuse*. We observe clear differences in the models that explain the evolution of male and female researchers within the same archetype ($p < .01$). Specifically, women and men differ within an archetype (e.g., *Diverse*) in where they start, transition rates, and their research interests during mid-career. Gender differences also exist in awarded grant income ($p < .05$) within a stage of an archetype. Regardless of gender, we also find significant differences ($p < .001$) in grant income as researchers evolve to next *behavioral stages* within an archetype. In general, we observe that income variability is accompanied by a shift in the dominant research area of the academic. It is worthwhile to note that we did not observe significant differences in grant income across archetypes.

We have strong quantitative results with competing baselines for future prediction and perplexity. For future *behavior* prediction, the proposed G-HMM cluster model improves by 24% over the best performing baseline. Our model also exhibits lower perplexity than the baselines.

1 Introduction

In this paper, we develop models to understand how individuals evolve with experience in social networks. The problem is important: as individuals interact with each other, they gain in experience, and behavioral changes reflect the newfound experience. However, despite a significant focus on community discovery and their evolution in social networks, our understanding of individual evolution is limited ([28, 42] are some notable exceptions). Understanding evolutionary patterns, in general, is useful in a variety of applications: language evolution [11]; expertise evolution [28]; journey optimization in digital advertising platforms.

Our specific interest lies in understanding how academics change their research behavior with gain in research experience. In the academic community, authors'

research interests are influenced by other authors' directly (collaboration) or indirectly (related published research) and in general, by the current research trends in the community. Analyzing the evolution of academic behavior on the community level has attracted persistent interest; previous works studied the evolution of research themes of a particular scientific community [23] or multiple communities [6, 7]. On an individual level, evolutionary studies have looked at modeling career transitions [32], citation evolution [37] and productivity or collaboration trends [40]. On the other hand, in this work, we want to identify dominant patterns of *research interests* evolution common among academics across different subfields. Our work can help to answer questions like, Do academics focus on a single research area throughout their career or they venture in multiple areas as they gain experience? If they work on multiple areas, when does this shift usually happens? Are some evolutionary patterns preferred over the others? This data-driven analysis of long-term researchers can assist in providing improved career guidance to junior faculty. Moreover, it can help funding agencies identify researchers of particular evolutionary patterns that may need more assistance.

At the outset, discovering patterns of individual evolution appears to be a combinatorial problem: academics vary in not only the sub-field that they choose to start but also in subsequent areas of interest. Furthermore, their research interests may evolve at different rates. Despite variations in the chosen sub-field of an academic, and how academics can evolve, we observe regularities at different stages of their career. For instance, for an academic, transition through different stages— Ph.D. Student (focusing on a single research area), being an assistant professor (working on few highly related areas) to eventually post-tenure (multiple areas, interests in multidisciplinary collaborations, etc.)—mark changes in research behavior. These elementary behavioral evolutionary patterns are visible in almost all academic fields, suggesting that surface variations (i.e., area of research for an academic) hide deeper regularities in patterns of behavioral change. We refer to these *latent* regularities in individual behavior as *behavioral stages*. We refer to the dominant progression patterns through behavioral stages as *archetypes*. Note that researchers may evolve at different rates through these stages. We show that we can explain all individuals' surface variations (the observed research area on which the academic focuses) with a small set of such archetypes. Fig 1 shows a stylized example.

Thus a model for learning archetypes needs to: express variation in observable research behavior while exhibiting latent stochastic regularities governing the change of behavior. Furthermore, the model should allow individuals to evolve at different rates. Finally, the results ought to be interpretable in a post-hoc manner.

Fig 1. A stylized academic evolutionary trajectory.
Each pie chart is a *behavior stage* in the trajectory. The numbers in each pie-chart show the fraction of papers published in each research area D_m in that stage. We use a normalized representation focused on the change of areas: the label D_1 represents the first research area of every academic, D_2 the second research area, etc. Normalized representations allow us to discover commonalities in behavioral changes of academics across seemingly unconnected domains. In this example, the top group of researchers evolves to shift their research focus to a new domain while the bottom group becomes increasingly interdisciplinary.

Our work makes the following contributions:

A framework for modeling evolutionary trajectories: We propose a sophisticated framework to identify dominant, interpretable, evolutionary archetypes amongst academics for modeling the evolution of their research interests. In contrast, prior work on academics has either focused on the

qualitative analysis (e.g., [38]) or predicting career transitions [32]. In our work, we assume that an archetype is a *probabilistic model* that encodes individual progression through stages of *distinct* behavior. Specifically, we learn a Gaussian Hidden Markov Model (G-HMM) to capture this progression where latent states capture *behavioral stages* in the evolution. To encode the idea of experience, while we allow individuals to evolve into the higher stages, we constrain our model to prevent individuals from returning to a stage from which they have evolved. We model *all* individuals with a *small* set of archetypes. We jointly learn the mapping of users into their archetype and the archetype's associated model's parameters through an Expectation-Maximization framework.

Finding: Dominant archetypes: While our framework is generic, we apply our model to understand the evolution of the research interests of Computer Scientists in this paper. We identify four archetypes with identical distribution of academics in our dataset: (i) *Steady* researchers who primarily work in their first research area throughout their career (most popular); (ii) *Evolving* researchers, who continuously shift their dominant area of research; (iii) researchers with *Diverse* research interests; and (iv) researchers who have *Diffused* interests with infrequent contributions in multiple areas. Each archetype is significantly different ($p < .001$) from the others.

Finding: variation by gender within archetype: We examine empirically, a subset of our data—all full professors (as of Spring 2018) in the top 50 CS departments in the United States for gender differences in their academic trajectory. We observe similar gender distribution across archetypes with the least number of women professors in *Evolving* archetypes. Moreover, within the same archetype, we observe significant differences in the models that explain the evolution of male and female researchers after inferring their trajectories individually. For instance, for the models that explain women and men differ ($p < .01$) in the *diverse* archetype; we observe men tend to start from later stages; 30% men (8% women) while women skip more stages; 50% women (36% men) skip stage 3; 14% women (9 % men) skip stage 2. Women also spend around a year more *exploring* mid-career than men (6.5 years for women vs 5.3 years for men) in the same stage.

Finding: variation in grant income: Next, we examine grant income (as of Spring 2018) from the National Science Foundation in the US for the same subset of CS academics, to understand the relationship between variations in awarded grant income over the course of academic trajectory and how difference in archetype or gender could serve as explanations. Although we did not observe any significant changes in the average grant income *between* archetypes, there exists income variability *within stages of each archetype*. In general, researchers are awarded more grant money as they gain experience, with the most notable uptick being after the first few years of their research career (between stages 2 and 3, $p < .001$) and in their last career stage (stage 5, $p < .05$). Specifically, researchers with *diverse* research interests receive subsequently increasing grant income while *evolving* researchers (who change their dominant research area in each stage) experience grant income variability with an area change. We find significant differences in grant income across genders *within a behavioral stage* of an archetype, mostly accompanied with stages marking a shift in dominant research area. For the *steady* and *diverse* archetype, female professors are awarded lower grant income than their male counterparts ($p < .05$) in their early career stages. On the other hand, evolving women receive a significantly lower income than

evolving men when they switch to new areas later in their career in stage 4 and 5($p < .05$).

Also, we have strong quantitative results with competing baselines for behavior prediction and perplexity on the Academic dataset. The proposed G-HMM cluster model improves by 24% over the baselines for behavior prediction. Our model also exhibits lower perplexity than the baselines.

Significance: We propose a sophisticated probabilistic framework to identify dominant, interpretable, evolutionary archetypes. We show that the discovered archetypes are significantly different and are straightforward to use to test hypotheses (e.g., evolutionary variation with gender; effects of gender on income). The framework is generic and is easily applied to model individual evolution in other social network datasets (e.g., Stack Exchange). Interested readers can refer to our arXiv submission [29].

2 Related Work

Prior literature related to our work is categorized into three broad categories: mining academia data, modeling time series, specifically using HMM models and studying the evolution of community and user in social network.

Academic Data Mining: There has been considerable interest in mining Academic Data (bibliographic data, researchers' usage of social media, etc.). Prior studies have looked at the evolution of research interests on a community level. Liu et al. [24] studied the evolution of research themes in articles published in CHI conference on Human Computer Interaction through co-word analysis. They highlighted specific topics as popular, core, or backbone research topics within the community. While Biryukov and Dong [6] compared different scientific communities in DBLP dataset in terms of its interdisciplinary nature, publication rates, and collaboration trends. They also studied the variation of author's productivity with career length and observed that most of the authors have a short career spanning less than five years. Chakraborty and Nandi [8] studied trajectories of successful papers in computer science and physics by analyzing paper citation counts. They classified these trajectories into multiple categories including early riser, a late riser, steady riser, and steady dropper.

Most of the work on trajectory mining on an individual level concerns career movement within academia. Deville et al. [13] observed that transitions between academic institutions are influenced by career stage and geographical proximity. While Clauset et al. [9] found that academic prestige correlates with higher productivity and better faculty placement. Recently, Safavi et al. [32] studied career transitions across academia, government, and industry for Computer Science researchers. Wang et al. [37] proposed a statistical model to predict the most impactful paper, in terms of citations, of scientists across disciplines. They argued nonexistence of a universal pattern and showed that highest-impact work in a scientist's career is randomly distributed within her body of work. Contrary to these studies, our focus is on finding patterns of change in *research interests* of scientists with gain in experience.

Recent studies also looked at gender differences in funding patterns, productivity, and collaboration trends in academia [39, 40]. Way et al. [39] did not observe any significant difference across gender in hiring outcomes in academia. However, they showed that indirect gender differences exist in terms of productivity, postdoctoral training rates, and in career growth. Some earlier studies also reported gender differences in academia. Kahn [18] identified gendered barriers in obtaining tenure for academics in economics, while Ward [38] found gendered differences in pay related to publication record. On the other hand, we explore *gender differences* in a

complementary dimension of diversity in research interests and its effect on awarded grant income.

Activity Modeling: Our work is most similar to activity sequence modeling that predicts next action or event in a sequence. We are different from those works as our focus is on modeling user *behavior* that is how a user spends her time among possible actions in each session. Yang et al. [43] and Knab et al. [19] proposed generative models that assign each action to a progression stage and classify event sequences simultaneously. They used their model to predict cancer symptoms, or products user would review in the future. However, the model did little to provide meaningful and interpretable stages and clusters. The major contribution of our work is in giving an *interpretable* model that helps us to characterize the temporal changes in user *behavior*.

Hidden Markov Model (HMM) has been used to model and cluster time sequences [5, 10, 35] in the past. However, most of these models learn an HMM for each user sequence and then employ clustering algorithms to cluster the learned HMMs. These approaches are not scalable, and the clusters thus identified are not interpretable. On the other hand, our model jointly clusters and learns model parameters resulting in much lower memory requirements.

User Profiling: There has been complementary work in the past on identifying and characterizing user roles in online social networks (OSNs). Maia et al. [26] identified five distinct user behaviors of YouTube users based on their individual and social attributes. While Mamykina et al. [27] identified user roles based on just answer frequency in StackExchange. Similar user behavioral studies are done by Adamic et al. [2] and Furtado et al. [14] on Yahoo Answers and Stack Overflow, respectively. These studies, however, ignore *temporal changes* in the behavior and use engineered features for behavior modeling.

Some behavioral studies do model the evolution too. Benevenuto et al. [4] learned a Markov model to examine transition behavior of users between different activities in Orkut in a static snapshot. Angeletou et al. [3] constructed handcrafted rules to identify user roles and study the change of user roles' composition in the community over time. Recently, Santos et al. [33] identified four distinct types of user activity pattern based on a set of activity-based features. Our model, instead, works directly on *raw activity data* and cluster users with a similar pattern of behavioral *evolution*. Although we have not evaluated on OSNs in this paper, our model has a generic framework and can be easily extended to these datasets as shown in [29].

3 Modeling Evolution Trajectories

In this section, we first introduce our dataset and then formally describe the problem statement followed by our proposed approach.

3.1 Data Collection

We use the Microsoft Academic dataset [34] provided through their Knowledge Service API (`http://bit.ly/microsoft-data`) to study evolutionary patterns of researchers with a focus on Computer Scientists. Microsoft Academic Service additionally annotates each publication with the year of publication, publication venue and the CS subfield (out of 35 identified fields) to which it belongs.

We can only query an individual author's publication history through the Microsoft Academic API. Thus, we create an author list to query by identifying *prominent* scientists from each of the 35 CS subfields. This comprehensive author dataset will help us discover dominant evolutionary archetypes common across authors from different

subfields. Also, *prominent* scientists usually have a long academic career to notice a considerable change in research interests.

We identify *prominent* authors based on *prestige* of the conference venues in which they publish, in their respective subfield. We use the older dump of Microsoft Academic dataset (`https://aminer.org/open-academic-graph`) to identify *prestigious* conferences for each subfield. Note that this older dump is noisy and contained multiple entries for the same authors; however, the online dataset is updated weekly, and API provides the most recent version. Thus, we decide to extract the updated author information from the API. We construct a conference-conference citation graph where each conference in our dataset forms a node, and the weighted edges represent inter-conference citation frequency. Specifically, the weight of a directed edge from conference C_1 to conference C_2 is proportional to the fraction of papers published in C_2 cited by papers published in C_1. We then use the Pagerank algorithm [1] on this directed graph and define conference *prestige* as the Pagerank of the corresponding conference-node. After that, we define an author's *prominence* as the weighted sum of the prestige of the conferences (s)he has published in. Here, conference-prestige are further weighted by the fraction of the author's papers published in that venue.

We rank authors in decreasing order of their *prominence* in each of the 35 CS areas (as annotated by Microsoft API) in the dataset. To get equal representation from all subfields, we then extract the publication history of top 750 most-prominent authors from each of the subfields in the dataset. Note that authors can be *prominent* in more than one subfield. We then filter unique authors from this set who have at least 15 years of publication history. This filtering is done to get a sufficient span of publication data to notice evolution in research interests. Further, we restrict our analysis to papers published from 1970 to 2016 to avoid missing data. The resulting dataset consists of records of 4578 authors with an average publication history of 24.15 years (This data will be made available upon publication).

3.2 Problem Definition

We represent an author's academic life-cycle as a sequence, $\mathbf{X_i}$, comprising of session-vectors, $\vec{X}_{ij}$. We keep *session* as a year-long since most conferences occur annually. Thus, $\mathbf{X_i}$ is a sequence of session-vectors, $\vec{X}_{ij}$, where $j \in \{1, 2, \dots t_i\}$ and t_i is the number of sessions for an author i. In general, lengths of sequences will vary across authors depending on the length of their academic career. A session, $\vec{X}_{ij}$, is a vector $\langle o_1, o_2, \dots, o_M \rangle$, where M denotes number of *area-of-interests* (`AoI`s). Each element o_m of the vector $\vec{X}_{ij}$, denotes the fraction of papers published in the D_m `AoI` by the i-th researcher during a single j-th year. This distribution of research areas of author's publications captures the research *behavior* of the individual in the year.

For defining an `AoI` of an author, we consider all papers published by the author in her academic life. We identify her primary `AoI`, D_1, as the *first* subfield (out of 35 subfields) in which she publishes *cumulatively* at least 3 papers in the first 3 years. Usually, an author's D_1 is about their Ph.D. dissertation work, and we expect students to *settle* down after a few years. Thus, after identification of D_1, hopefully with a steady paper count, we define her secondary `AoI`, D_2, as the subfield in which she publishes at least 3 papers in *one* year. Similarly, we also define tertiary (D_3), quaternary (D_4), and quinary (D_5) `AoI`. We do not define `AoI`s beyond D_5 because 80% of authors do not explore more than 5 subfields in our dataset. Also, in a given year, if an author publishes fewer than 3 papers in an unexplored subfield, these papers count towards a sixth dimension `AoI` called *Explore* (Ex). *Explore* dimension denotes that the author has started exploring new subfields but are not notable enough to be one of the D_m's ($m \in [1, 5]$), and indicate a possible shift in research interests.

To summarize, each session is a 6 dimensional vector ($M = 6$), and its elements are fraction of the author's publications in the 5 D_m's or the 6^{th} *Explore* dimension. This normalized session representation allows our model to discover behavioral patterns of the author's changing research interests in a domain-independent manner. For example, in a given year, the session-vector for an author who publishes 3 papers in theory (D_1; primary area) and 1 paper in graphics (D_2; secondary area), and the session-vector for another author who publishes 3 and 1 papers in NLP (D_1; primary area) and ML (D_2; secondary area) respectively will be exactly same: $X_{ij} = \langle 0.75, 0.25, 0, 0, 0, 0 \rangle$. Notice that normalization does not change the rate at which a specific author decides to switch domains and is also invariant to subarea publication norms ([39] observed productivity rates differ by subfield in DBLP).

The problem then addressed in this paper is to associate an *archetype* with each author's sequence. We assume that there exist C different archetypes, and given a sequence of session-vectors for an author $\vec{X}_i = \{\vec{X}_{i1}, \ldots \vec{X}_{it_i}\}$, the goal is to assign the sequence to one of the C *archetypes*—each associated with a set of K latent *behavioral stages*. During this assignment, we also identify how the individual evolves through its archetype's distinct stages by outputting the sequence $Y_i = \{Y_{i1}, Y_{i2} \ldots Y_{it_i}\}$, where Y_{ij} represents the behavioral stage $k \in [1, K]$ assigned to j-th session in individual i's sequence. We constrain the number of stages $K \ll t_i$ and allow skipping of stages while disallowing return to earlier stages.

3.3 A Framework for Identifying Archetypes

We use a Gaussian-Hidden Markov Model (G-HMM) based approach to model individual behavior. In our model, latent states of the G-HMM capture the *stochastic regularities* in behavior while Gaussian observations enable *variations* in the session-vector distributions (instead of fixed observations in vanilla HMM). Thus, a G-HMM captures an archetype with all individuals belonging to the archetype, going through the same set of *behavioral stages* or latent state. Note that G-HMM allows for skipping states and variable evolutionary rates among individuals.

To capture broad variations amongst individuals, we learn a set of C G-HMMs where each G-HMM represents a distinct archetype. We jointly learn the partitioning of the individuals into different archetypes and the model parameters for each archetype.

Each Gaussian HMM, associated with an archetype c, has K discrete latent states or *behavioral stages*. The model makes a first-order Markovian assumption between state transitions using the transition probability matrix $\tau^{\mathbf{c}}$; where τ^c_{kl} represents the probability of transitioning from stage k to l in the c-th archetype. The prior probabilities of the latent states are represented by the K dimensional vector π^c. Lastly, the model assumes that given a latent behavioral stage, k, from an archetype c, the M dimensional session vector, X_{ij}, is Normally distributed with mean μ^c_k and covariance $\mathbf{\Sigma^c_k}$. The mean vector μ^c_k essentially encapsulates the typical behavior exhibited in the k-th *behavioral stage*.

In the above model, the G-HMM associated with different archetypes do not share latent states. In other words, each G-HMM has its own set of discrete latent states (Experiments with tied-states of archetypes led to worse results.). However, we fix the number of states (K) to be the same for each archetype.

Encoding Experience & Variable Evolutionary Rates: To encode the idea of experience, as well as to allow variable evolutionary rates, similar to [43] and [19], we allow only forward state transitions (including self-loop) within a G-HMM that represents an archetype. This choice appears sensible to us since semantically, each latent state of the G-HMM represents a *behavioral stage* of evolution, and its corresponding mean vector encapsulates *behavior* in that stage. Then, forward

transition captures *progression* through *behavioral stages*. We operationalize this idea by restricting the state transition matrix to be an upper triangular state transition matrix.

Algorithm 1: Gaussian HMM archetype

Input: $\vec{X_i}$ and $\lambda_{\mathbf{0}}^{\mathbf{c}}$ $\forall i \in \{1, 2, \dots N\}$ $\forall c \in \{1, 2, \dots C\}$;
Output: $\vec{Y_i}$ and $\lambda^{\mathbf{c}}$ $\forall i \in \{1, 2, \dots N\}$ $\forall c \in \{1, 2, \dots C\}$;
Initialize the c^{th} archetype with initial parameters, $\lambda_{\mathbf{0}}^{\mathbf{c}}$ $\forall c$;
while *not converged* **do**
  M-Step: Re-assign archetypes to sequences $\mathbf{X_i}$ as:
  $c_i = argmax_c P(\mathbf{X_i}|\lambda^{\mathbf{c}})$ $\forall i \in \{1, 2, \dots N\}$;
  E-Step: Re-estimate the G-HMM parameters, $\lambda^{\mathbf{c}} \forall c \in \{1, 2, \dots C\}$, using modified Baum-Welch algorithm.;
end
Convergence Criteria;

- Log Likelihood difference falls below threshold; or
- Number of iterations is greater than threshold; or
- Number of sequences re-assigned in an iteration is less than 1% of the data

Training: We train our G-HMM cluster model using a (hard) Expectation Maximization [12] based iterative procedure described in Algorithm 1. During training, the goal is to learn the G-HMM parameters, $\lambda^{\mathbf{c}}$, for each archetype c, where $\lambda^{\mathbf{c}} = \langle \mu^c, \mathbf{\Sigma^c}, \pi^c, \tau^c \rangle$ and archetype assignments for each user, c_i. We first initialize the Gaussian HMMs with initial parameters, $\lambda_{\mathbf{0}}^{\mathbf{1}}, \lambda_{\mathbf{0}}^{\mathbf{2}}, \dots, \lambda_{\mathbf{0}}^{\mathbf{C}}$. After that, in the iterative training process, in the Expectation step, we use current estimates of $\lambda^{\mathbf{c}}$ to assign an archetype to each user sequence in the data. In the Maximization step, we use current archetype assignments to learn the corresponding G-HMM's parameters, $\lambda^{\mathbf{c}}$. We use a modified version of the Baum-Welch algorithm [30], allowing for forward-only transitions. Thus, this method jointly partitions the input sequences into different archetypes as well as learns the parameters of the associated G-HMMs.

Implementation Details: Our iterative training procedure requires initialization for G-HMM parameters, $\lambda_{\mathbf{0}}^{\mathbf{c}}$. We perform k-means clustering on all sessions of all user sequences in our corpus, treating the sessions as independent of each other (thus losing the sequential information). The cluster centers, thus obtained are used as the initial means, μ_0^c, for the latent states. We fix each Σ_k^c as an identical diagonal covariance matrix σI with $\sigma = 0.01$ based on preliminary experiments. We initialize transition matrices, τ_0^c, and states' prior probabilities, π_0^c, for each archetype randomly. Our implementation is based on Kevin Murphy's HMM Matlab toolbox(`bit.ly/hmmtoolbox`). Also, we implement a parallelized version of our EM algorithm to reduce computation time. We test our model on Intel Xeon Processor with 128 Gb RAM and a clock speed of 2.5 GHz.

4 Result Analysis

In this section, we perform analysis of archetypes identified by our model. We first describe the discovered archetypes of all researchers in Section 4.1. Then, we examine gender variation in academic trajectory in Section 4.2 and effect of archetype and gender on grant income in Section 4.3.

Behavioral Stage	Steady	Diverse	Evolving	Diffuse
Stage **1**	{3Y, 5m} **D_1** (87%) Ex (11%)	{3Y, 3m} **D_1** (88%) Ex (11%)	{2Y, 9m} **D_1** (72%) Ex (24%)	{2Y, 7m} **D_1**(76%) Ex (22%)
Stage **2**	{4Y, 2m} **Ex** (74%) D_1 (23%)	{2Y, 6m } **Ex** (80%) D_1 (16%)	{2Y, 9m } **Ex** (83%) D_1 (12%)	{3Y, 7m } **Ex** (91%)
Stage **3**	{7Y, 5m} **D_1** (62%) **Ex** (32%)	{5Y, 6m} **D_1** (73%) Ex (17%)	{6Y, 2m} **D_1** (33%) **Ex** (28%) **D_2** (24%)	{8Y, 5m} **D_1** (50%) **Ex** (39%)
Stage **4**	{5Y, 9m} **D_2** (49%) **D_1** (27%) Ex (17%)	{5Y, 6m} **Ex** (46%) D_2 (20%) D_1 (17%)	{5Y} **D_2** (66%) *Ex* (18%)	{3Y, 9m} **D_2** (43%) **Ex** (26%)
Stage **5**	{2Y, 6m} **D_1** (49%) Ex (18%) D_2 (14%)	{6Y, 3m} **D_4** (29%) **Ex** (20%) D_3 (14%) D_1 (14%)	{6Y, 5m} **D_3** (43%) Ex (19%) D_2 (14%)	{4Y, 1m} **Ex** (74%)

Table 1. Learned mean vector for each latent state of four archetypes in the Academic Dataset. We list the *Area-of-Interests* (`AoI`) in sorted order and annotate them with their % contribution in the state. We only list significant `AoI` (> 11%) for each state. Each state is also labeled with its average duration in {Years (Y), months (m)}. The labels given to these clusters reflect our interpretation of the user behavior and make disambiguation of the behavior easier in the text.

4.1 Discovered Archetypes

Our analysis reveals four dominant archetypes: *Steady*, *Diverse*, *Evolving* and *Diffuse*. We chose the number of clusters $C = 4$ using the elbow method [36]: data log-likelihoods increased rapidly till four clusters with much slower increase beyond that. Further, we chose the number of states per cluster, $K = 5$: beyond five states, KL divergence[21] between mean vectors of new states with previous states started reducing rapidly, indicating redundant states.

We also conducted t-test to validate differences among the identified archetypes. Specifically, paired-sample t-test [16] is conducted between likelihood values of data points assigned to an archetype with their likelihood values obtained from rest of the archetypes. For instance, for each archetype pair (p, q), we conduct paired t-test between $\log P(X_i|\lambda^p)$ and $\log P(X_i|\lambda^q)$ $\forall i \ni c_i = p$. Note that test results for archetype pair (p, q) are not symmetric. We observed that all archetype pairs are significantly different ($p < .001$) from each other. Now, we proceed to discuss what is common to these discovered archetypes before examining each one in detail.

`Commonalities in Archetypes:` Table 1 summarizes the trajectories (state sequences) learned for the four different archetypes in this dataset. Each archetype is labeled according to our interpretation of the user behavior, looking at the learned mean vector of G-HMM states. We observe that all archetypes exhibit similarities,

especially in the first two stages. Across all archetypes, the first *stage* typically spans around 3 years, and more than 72% of the published research is in the author's *primary* `AoI`: D_1. As noted before, this is most likely their Ph.D. dissertation area, and hence, the research is more focused. After gaining some research experience, most authors move to the second *stage* where they start exploring other research areas denoted by a marked increase in their *Explore* `AoI`(more than 74%). However, in state 3 and beyond, authors from different archetypes follow different trajectories where they differ in how they change their dominant `AoI` over time while *exploring* other domains. We also observe that all archetypes have similar average career length and author count (Table 1 and Table 2). Below, we describe each archetype in more detail.

`Steady`: The first major archetype is of *steady* researchers, who mainly work in *one* `AoI` (i.e. their D_1) throughout their career. Fig 2 shows the state sequence of this archetype. We can see that most people start in their primary `AoI`, D_1 (state 1), which possibly reflects their Ph.D. education. After graduation, they spend some time *exploring* other areas while continuing to publish in D_1 (state 2), but move back to publishing in D_1 for a significant portion of their careers, about 7.5 years (state 3). This shift is often again followed by a phase where they start working in another area, D_2, while continuing to publish in D_1 (state 4). They eventually revert to publishing in D_1 (state 5) towards the latter part of their careers. In the last state, they also publish widely in other areas (indicated by almost half of the pie divided between other D_m's), but their main interest remains D_1.

Fig 2. Trajectory (state sequence) for *Steady* archetype in the Academic Dataset. Each pie is a *latent state* or *behavioral stage* in the trajectory. It denotes the mean proportion of papers published in each *Area of Interest*'s in the latent state. Each state is also labeled with the average amount of time spent in the state. For example, in this cluster, 87% of publications in the first 3.5 years are in the author's primary `AoI` D_1 while rest 11% are in exploring other areas. The arrows on the top of each pie show the prior probability for starting in that state. As we learn a left-to-right G-HMM, an author can transition to its immediate next state or any later latent states. Each transition is labeled with the corresponding conditional transition probability i.e., transition probability given that the user has decided to transition. The arrows thickness is proportional to its weight. Authors in this cluster exhibit *steady* research interest in their primary `AoI` D_1. Some authors start contributing dominantly in their secondary `AoI`, D_2 in State 4. Though, they return to spending around half of their effort in D_1 in State 5.

For example, Michael Jordan, professor at the University of California, Berkeley exhibits this research trajectory. He is a Machine Learning expert; his primary `AoI` D_1, and has secondary interests in Data Mining, Optimization, and Bioinformatics. Theory professor at University of Illinois, Urbana-Champaign (UIUC), Jeff Erickson is also assigned to this cluster; he also publishes in his primary `AoI` D_1 (Theory) with auxiliary interests in mathematical optimization.

`Diverse`: The second archetype consists of researchers with *diverse* research interests as they make significant contributions in multiple D_m's. Similar to *steady* researchers, these researchers research in their primary `AoI` D_1 while *exploring* other domains in the initial 3 states as shown in Table 1. They, then, publish in D_2 and D_1 while spending half time *exploring* other possible interests (state 4). They evolve to have a strong research presence in all 5 AoIs (state 5). This behavior suggests that authors of this archetype tend to work in interdisciplinary areas; or possibly projects with a broader scope which gains acceptance by different research communities. One notable example is Prof. Jiawei Han at UIUC, who started his academic career studying Databases and Data Mining, is also making notable contributions in Machine Learning and

Bioinformatics lately. Another professor who started in Databases, Jaideep Srivastava of the University of Maryland, evolved on to research distributed implementation of databases, and also data mining and AI-related research simultaneously.

`Evolving`: These researchers have one dominant area of interest (`AoI`) in each state which *changes* with time. Their dominant area of interest (`AoI`) *evolves* from D_1 (72%) in state 1 to D_2 (66%) in state 4 to D_3(43%) in state 5. Even though their `AoI` shifts across stages, in any given stage, they remain focused on one area and do not publish much in other areas. James Foley, a professor in Georgia Tech, started in Computer Graphics and later switched to research on user-computer interfaces and recently, User Modeling. Natural Language Processing (NLP) expert Daniel Jurafsky at Stanford University, also steadily moved from pure NLP based research problems to Speech processing, and later to Machine Learning (ML). Also note, for Jurafsky, this evolution can be attributed to the broader field shift of using sophisticated ML models to solve NLP problems.

`Diffuse`: Authors of this archetype stay focused in one dominant area in each stage; while in the last stage, their research interests are *diffused*. Authors publish considerably in one dominant area in first 3 stages; D_1 (state 1, 3) to D_2 (state 4). In the last state, which lasts around four years, the authors are infrequently publishing (less than three papers a year) in new subfields accounting for 74% of their publications. Hence, these authors have *diffused* research interests after they gain experience. Gerhard Weikum, professor at MPI Germany started in Databases area made a brief transition to Information Retrieval work and later started publishing in Machine Learning and Data Mining fields too. These area evolutions seem to be natural transitions as they are highly interrelated, which explains contributions in all fields. Anind Dey, professor at Carnegie Mellon University, initially worked on sensor technology and then switched to Web mining and Human Computing related research problems is also another example of this archetype. Dan Cosley, a professor in Cornell specializing in Information Science, is also another great example who publishes consistently in his D_1 (World Wide Web) and D_2 (Knowledge management) and lately in Data Mining.

4.2 Archetype variations across Gender

We now proceed to analyze the variations in the evolution of research interests (or archetypes) between male and female researchers. To this end, we manually annotate gender of all current and emeritus professors in top 50 Computer Science (CS) Universities as reported by U.S. News & World Report (`bit.ly/usnews-cs`). We consider only current and emeritus *Full* Professors as they typically have 15 or more years of publication history. This results in a total of 1084 authors in our dataset, 127 of whom are women. Table 2 shows the distribution across archetypes. We observe similar gender distribution in each archetype with the least number of women academics in *evolving* archetype.

Social Group	Steady	Diverse	Evolving	Diffuse
Male Professors (Top-50 US schools)	247	206	241	263
Female Professors (Top-50 US schools)	30	32	26	39
All authors in the dataset	1329	1080	1107	1062

Table 2. Statistics for discovered archetypes in relationship to different social groups.

While researchers from both genders in the same archetype c will traverse the same set of stages, they may differ in *how* they transition τ^c, and at *which* stage they start π^c. For this analysis, we first run our model on the entire dataset assigning archetypes

to each individual. We then estimate separate model parameters for female λ_f^c and male λ_m^c researchers for each archetype c using the assigned values.

To quantify the difference between two models $(\lambda_f^c, \lambda_m^c)$ for archetype c, we compute their *likelihood ratio*. Likelihood ratio R_f^c of female researchers in archetype c is:

$$R_f^c = \exp\left(\frac{1}{|N_f^c|}\sum_{i \in N_f} \log \frac{P(X_i|\lambda_f^c)}{P(X_i|\lambda_m^c)}\right) \quad (1)$$

where N_f^c represents all female researchers in c-th archetype. The equation simplifies to say that $\log R_f^c$ is the average difference between log-likelihoods of a trajectory of a female researcher generated from their own model with those of male model of the same archetype. Thus, for instance, value of $R_f^c = 2$ denotes that female researchers are twice more likely to be generated by the model of their own gender than of the opposite gender. We compute a similar ratio, R_m^c, for men. We also conduct paired-sample t-test [16] between the two likelihood values similar to Section 4.1.

Table 3 shows the likelihood ratio with their p-values. Since most of the values are statistically significant, all researchers are better explained by the model for their gender, than by the model for the opposite gender. Male researchers are distinct for the steady and diverse archetypes, but not for the evolving and diffuse archetypes. For women, on average, the difference is larger, with the strongest difference seen for the steady, diverse, and evolving archetypes.

Gender	Steady	Diverse	Evolving	Diffuse
Male	2.10***	2.63**	1.15	1.10
Female	1.80***	1.64**	1.60***	1.38***

Table 3. Likelihood ratio for academics across genders within an archetype. It measures odds of a researcher being better explained by model for their gender than by model for the other gender.
$* = p < .05, ** = p < .01, *** = p < .001$

For the sake of brevity, we examine gender difference in only the *diverse* archetype in some detail. Fig 3 shows three interesting variations. First, we observe that women are much more likely to start in state 1 (92%), with a dominant area of interest (D_1) than in any other state. In contrast, men start in states 1, 2, 3, and 4, with only 70% starting in state 1. Both men and women skip stages, but women are more likely to skip a stage than men. For example, 50% of women skip stage 3, while only 36% of men do. Longer skips of two stages are rarer, and both women and men make these long skips at the same rate. Finally, there are clear differences between mid-career men and women (states 3, 4): women spend more time *exploring* mid-career (state 4) than men, and mid-career men spend more time in their starting area of interest (D_1, state 3) than women.

Fig 3. Gender wise representation of trajectory for researchers belonging to the *diverse* archetype in the Academic Dataset. The transitions in blue denote transition probabilities of female professors in the archetype while those in red represents probabilities for their male counterparts. Men start their career from later evolved stages while women make long term state transitions.

4.3 Grant income variability across Archetypes & gender within an Archetype

We next examine the relationship between variation in the academic trajectories and gender to research grants awarded at different stages of an academic career. We extract historical information of grants from the National Science Foundation, a large federal funding agency for Science & Engineering in the United States (`bit.ly/nsfgrants`). We consider grants with Principal Investigators (PI) from the same subset of CS professors in top-50 US universities as in Section 4.2. We collect information for 1062 professors and manually disambiguate names and identify gender by cross-validating with the researcher's webpage. Then, we compute the average grant money awarded to a researcher, at each stage in their trajectory. Fig 4, which shows letter-value plots of average grant size awarded as PI's, broken down by archetypes (steady, diverse, evolving or diffuse), stage within an archetype and gender, summarizes our findings.

Fig 4. Letter value plots of total grant money awarded by NSF when author is a PI in each stage. a) Steady Researchers, b) Diverse Researchers, c) Evolving Researchers, d) Diffusive Researchers. In general, Professors get more grant money as they gain experience. Regardless of archetypes, grant income in state 3 is significantly higher from state 2 ($p < .01$). There are also significant differences across genders within a state of an archetype. For instance, for Evolving archetype, male professors get significantly more income than female professors in state 4 ($p < .01$).

We conducted Kruskal-Wallis H-test [20] to establish difference between average grant income awarded in the same state across archetypes. However, this test failed indicating similar average income of researchers from different archetypes while in the same career stage (Fig 4). Consequently, we conducted H-test to establish the statistical significance of differences in grant money across latent states within an archetype. Success of this test affirmed that at least one latent state is different from another latent state within an archetype. We then conducted Welch's t-test [41] between consecutive states to find the exact pair of states which are significantly different. We only tested with consecutive latent states as we are only interested in grant income changes as the author progresses through stages. Table 4 reports the state pairs for each archetype that are statistically different. In the rest of this section, we describe these results in detail.

Archetype (H-test)	State Pair (t-test)
Steady***	State 2 vs 3**
Diverse***	State 2 vs 3***
	State 4 vs 5*
Evolving***	State 2 vs 3***
	State 3 vs 4*
	State 4 vs 5**
Diffuse***	State 2 vs 3***
	State 4 vs 5*

Table 4. Statistical significance tests for the differences in grant money across latent states within an archetype. Shown are only those tests that are statistically significant. H-test [20] confirms that at least one state is different from another state of the archetype; t-test [41] was then conducted between each consecutive states within the archetype to determine the differing states.
$* = p < .05, ** = p < .01, *** = p < .001$

Regardless of archetypes, we observe that in general authors tend to receive more

grant money as they gain experience in Fig 4. On average, across archetypes and gender, PI's receive in state 5, four times the amount of grant money than state 1 ($p < .001$). Also for researchers across archetypes and across genders, we notice an uptick in grant income in state 3 from state 2 ($p < .01$ - Table 4). Let us qualitatively examine the *steady* researchers in detail, by comparing Fig 4 with Fig 2. State 2 in Fig 2 shows the researchers exploring different topics, whereas, in state 3, they are spending a significant part of their time on their main domain D_1. Also, notice that 36% of the researchers never visit state 2 - 27% skip state 2, and 9% of the researchers start in state 3. Since state 1 typically represents the time spent by the researchers in their Ph.D., and with 74% time spent in an explore stage in state 2, it is not surprising that we see limited grant income in their first two states. State 3, perhaps reflects a sustained focus on their domain D_1, and this pays off in terms of grant income. Similar qualitative arguments follow for the other archetypes.

However, the grant trajectories over states is different for each archetype ($p < .001$). Let us examine statistically different state pairs from Table 4 in Figure 4. Steady researchers see a big uptick in their grant income in state 3 and their subsequent grant income is similar in magnitude ($p < .01$). The grant income for diverse researchers (who have more than one dominant area) increases steadily over states($p < .05$). For evolving researchers (who change their dominant area), the grant income rises (state 3, $p < .001$), falls (state 4, $p < .05$) and rises (state 5, $p < .01$), reflecting a degree of unpredictability accompanying *changing area of interest*. Diffuse researchers, have a pattern similar to steady researchers except in state 5 ($p < .05$), when the income dips, perhaps due to spending time in *too many areas*.

To determine differences in grant income across gender, we further conducted t-test [41] between grant distributions of female and male professors in each state within an archetype. Table 5 reports significantly different states within each archetype. We again examine these statistically different states from Table 5 in Fig 4. Evolving women receive significantly *lower* income than evolving men when they *switch to new areas* in state 4 and 5 ($p < .05$). On the other hand, in our dataset, steady women receive *higher* grant income than steady men when they *switch areas* in state 4 ($p < .05$). In general, we observe that men show greater grant income variability than do women. The variability is statistically significant ($p < .05$) during early career in state 1 and 2 for Steady and Diverse researchers respectively. We do not observe significant differences in grant income of male and female Diffusive researchers.

Archetype	Latent State (t-test)
Steady	State 1**
	State 4**
Diverse	State 2**
Evolving	State 2*
	State 4***
	State 5**
Diffuse	Not significant

Table 5. Statistical significance tests [41] for the differences in grant money across gender in each state within an archetype. Shown are only those tests that are statistically significant.
$* = p < .1, ** = p < .05, *** = p < .01$

In summary, we identify four dominant archetypes for researchers: steady, diverse, evolving, and diffuse. We observe differences in the evolution of male and female researchers within the same archetype. When we examine the diverse archetype in detail, we observe that women and men differ in where they start, rate of transition,

and time spent in mid-career. The differences in grant income are salient across states within an archetype. In general, grant income increases with experience. We also observe differences across genders within a stage of an archetype. We noticed that grant income variability is mostly accompanied with stages marking change in dominant area of interest. Lack of difference in average grant income across archetypes indicate that different ways of conducting research do not have a significant impact on the income. However, researchers can expect variability in the income sporadically when venturing into new areas, especially for women professors.

5 Quantitative Experiments

In this section, we show effectiveness of our model by evaluating it on two different tasks: Future Prediction and Perplexity. We describe the baselines in Section 5.1 and report results in Section 5.2.

5.1 Baselines

Distance G-HMM: Our first baseline uses the G-HMM clustering model as defined in [15]. In this baseline, we learn a G-HMM for each user and then cluster the models using distance metric δ, the symmetric KL divergence (d_{kl}) between two G-HMMs [17].

$$d_{kl}(\lambda^p, \lambda^q) = \frac{1}{N_p} \sum_{i \in N_p} log \frac{P(X_i|\lambda^p)}{P(X_i|\lambda^q)}, \tag{2}$$

We use k-medoids clustering; since this method does not give a representative model for each cluster, we additionally learn a G-HMM per cluster. For a fair comparison, we set k, the number of clusters to be the same as our model.

Vector Autoregressive Model (VAR): VAR models are used to model multivariate time series data [25]. It assumes that each variable in the vector is a linear function of its own past values as well as other variables. For each user sequence $\mathbf{X_i}$, jth session is modeled as,

$$\vec{X}_{ij} = A_1 \vec{X}_{ij-1} + \ldots + A_p \vec{X}_{ij-p} + u_j \tag{3}$$

where A_i is M X M matrix, $u_j \sim \mathcal{N}(0, \Sigma_u)$ and we set $p = 1$ as in first-order Markov models.

No Evolution: In this baseline, we assume that individuals *do not evolve* in their lifespan. This baseline is a simplified version of our model. It assumes that there are different archetypes but that each archetype has only one state. Hence, all sessions of a sequence are generated from a single multivariate Gaussian.

Prior work on activity sequence prediction baselines [19, 43] deals with discrete data. However, as we represent each session as a continuous vector, these approaches are not directly comparable and adapting them to our problem is nontrivial.

5.2 Tasks

Future Activity Prediction: In this task, we predict the future behavior of an individual given her history. We assign the first 90% sessions of each sequence for training and predict the behavior in future sessions (the remaining 10% of the sequence). We first use all the training sessions to learn the parameters of our model. Then, for each sequence, we run the Viterbi algorithm to decode the state assignment of its test sessions, t'_i. The test sessions of the i-th user will have same archetype assignment c_i determined in the training session for that user.

We compute Jensen-Shannon(d_{js}) divergence between the mean $\mu^{c_{ij}}$ of the assigned state Y_{ij} and the observed vector X_{ij}. d_{js} is a symmetric K-L divergence between two vectors. We report the average $\bar{\Delta}$ over all test sessions:

$$\bar{\Delta} = \frac{1}{|T|} \sum_{i \in N, j \in t'_i} d_{js}(\mu^{c_{ij}}, X_{ij}), \tag{4}$$

$$d_{js}(\mu^{c_{ij}}, X_{ij}) = \frac{1}{2} d_{kl}(\mu^{c_{ij}}, p) + \frac{1}{2} d_{kl}(X_{ij}, p), \tag{5}$$

where, $p = \frac{1}{2}(\mu^{c_{ij}} + X_{ij})$ and d_{kl} measures KL divergence distance. For VAR, we use the model learnt on training sessions of user i to make prediction for her future sessions.

Table 6 shows our results on this task. Our model outperforms the baselines by 24% on the Academic dataset. It shows that learning archetypes can also help us to accurately predict an individual's future behavior in the social network.

Metric	Our Model	No Evolution	VAR[25]	Distance G-HMM[15]
Future Prediction	0.22	0.42	0.31	0.29
Perplexity	−18.37	100.79	NA	37.73

Table 6. First row lists the average Jensen-Shannon divergence of future sessions using 90-10% split of each user sequence. While second rows lists average perplexity on unseen user sequences after 5 fold cross validation. Lower values are better. Note negative log values are because of continuous densities.

Perplexity Perplexity measures how surprised the model is on observing an unseen user sequence. A lower value of perplexity indicates low surprise and hence a better model.

$$P_x = -\frac{1}{|T|} \sum_{i \in T} \max_{c \in C} \left(\log P(\mathbf{X_i^T} | \lambda^c) \right) \tag{6}$$

where, $\mathbf{X_i^T}$ represents a test sequence in Test Set T, and λ_c represents the parameters of the G-HMM corresponding to the c-th archetype. We assign $\mathbf{X_i^T}$ to the archetype c with maximum likelihood. Perplexity is then computed as the average likelihood of all test sequences. In general, $P(\mathbf{X_i^T} | \lambda^c))$ is bound between [0,1] but as we model continuous data with multivariate Gaussian distribution, probability is computed as a density function and can be > 1.

Table 6 also reports average perplexity after five-fold cross-validation. Note that for this experiment, the model predicts the entire trajectory of a new user. We could not use the regression baseline (VAR) as it is not a generative model and can not predict an entirely new sequence. Our model beats the best performing baseline by 149% on our dataset. Hence, our model also effectively predicts the behavior of future individuals joining the social network. Note that our model gives negative perplexity values i.e., negative log values. It indicates that the likelihood is more than one due to the Gaussian kernel, as mentioned earlier.

Discussion: For future prediction, our model performs better than the VAR model. It shows that modeling cluster of sequences gives a better estimate than modeling each user sequence separately. Also, if we assume no evolution and just cluster users according to their behavior i.e., *No Evolution* model, we obtain worse results indicating that individuals behavior does not stay constant over time.

Our model also outperforms the similarity distance-based clustering method: Distance G-HMM [15], which is also the strongest baseline. It first estimates the G-HMM model for each user sequence and then clusters these models. Estimating

model for each sequence can be noisy, especially if the user sequence has a short length. Instead, when we jointly learn G-HMM model parameters and cluster sequences, we learn a better approximation.

Full vs. Left-Right Transition Matrix: We also test our model with unconstrained full transition matrix where users can jump from one state to any other state in the HMM. We obtain slightly better results with this model for the future prediction task. This improvement can be due to more degrees of freedom, but then, it is also computationally expensive to learn. However, our model gives comparable results with much fewer parameters. Also, with full transition matrix, learned states are not interpretable in the context of evolution. As [43] and [19] also noted, forward state transitions accurately models the natural progression of evolution, we thus chose to work with a forward transition matrix.

6 Limitations and Future Work

Our proposed model identified insightful archetypes and its variability with gender and grant income of professors. However, it is essential to understand certain caveats to the reported findings. First, in terms of the data, the discovered archetypes for academics are for the top researchers in their field (we pick *prominent* researchers in each of the 35 research subdomains) with a long career span (15 years). Thus, our archetypes do not reflect junior scientists engaged in research or researchers with sporadic research output. Nonetheless, our study will help junior scientists to understand diverse ways of research behavior of successful academics in the field and tailor their career to their interests. In our current study, we collected grant history from public data available by NSF. The funding analysis can be extended by collecting data from other possible funding sources like the National Institute of Health (NIH), gifts from industry, and professor's salary. The findings of gender bias can be different if we include these private sources of income ([38] observed gender differences in professor's pay). Hence, we believe that our study is the first step in understanding the effect of research conducting behavior of academics on their income.

Second, as with all inductive models, our qualitative results depend on the chosen model. Recently, Deep Neural Networks, especially Recurrent Neural Networks, have been proposed to model time series data. There has also been considerable interest in building interpretable neural models [22, 31]. However, still neural approaches are hard to interpret, and developing interpretable neural prediction models is something that we plan to look at in future work.

Third, in our current version of the model, we do not consider the effect of collaborations or the role of conferences where researchers publish, and where they may pick up on normative behavior (e.g., areas in which to work) on the discovered archetypes. In future work, we plan to understand the role of community interaction on archetypes and address these limitations. Another interesting research direction is to explore the correlation of change in research behavior with career transitions and author's citation count.

7 Conclusion

In this paper, we aimed to discover the archetypical research behavior of Academics. The observation that despite surface variation in terms of sub-fields, the change in behavior exhibits regularities, motivated our research. We introduced a novel Gaussian Hidden Markov Model Cluster (G-HMM) to identify archetypes and evolutionary patterns within each archetype. We chose to work with G-HMM's since they allow for:

variations in trajectories and different evolutionary rates; constrain how individuals can evolve; are interpretable.

We identified four distinct archetypes of computer scientists: steady, diverse, evolving, and diffuse and showed examples of computer scientists from different sub-fields that share the same archetype. We analyzed full professors from the top 50 CS departments to understand gender differences within archetypes. Women and men differ within an archetype (e.g., diverse) in where they start, rate of transition and research interests during mid-career. We further analyzed grant income of these professors to understand the effect of gender and archetype on income. The differences in income are salient *across states* within an archetype rather than *across archetypes*. There also exist significant differences *across genders within a state* of an archetype. We observed that most of the grant income variability across states or gender is accompanied by a shift in the dominant research area of the academic. In light of our findings, we propose the funding agencies and academic institutions to be wary of these differences and be supportive of researchers venturing into new areas.

To the best of our knowledge, we are the first one to provide a principled framework to model and identify interpretable individual trajectories in academia. Our model can be easily used to identify trajectory in other domains like medicine, physics, and business. Further work on the comparison of research trajectories from the stem and non-stem fields can be an exciting research direction.

References

1. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, November 1999. URL `http://ilpubs.stanford.edu:8090/422/`.

2. Lada A. Adamic, Jun Zhang, Eytan Bakshy, and Mark S. Ackerman. Knowledge sharing and yahoo answers: Everyone knows something. WWW'08, 2008. ISBN 978-1-60558-085-2. doi: 10.1145/1367497.1367587. URL `http://doi.acm.org/10.1145/1367497.1367587`.

3. Sofia Angeletou, Matthew Rowe, and Harith Alani. Modelling and analysis of user behaviour in online communities. In *The Semantic Web-ISWC'11*, 2011. ISBN 978-3-642-25073-6.

4. Fabricio Benevenuto, Tiago Rodrigues, Meeyoung Cha, and Virgilio Almeida. Characterizing user behavior in online social networks. IMC'09, 2009. ISBN 978-1-60558-771-4. doi: 10.1145/1644893.1644900. URL `http://doi.acm.org/10.1145/1644893.1644900`.

5. Manuele Bicego, Vittorio Murino, and Mário A. T. Figueiredo. Similarity-based clustering of sequences using hidden markov models. In *Proceedings of the 3rd International Conference on Machine Learning and Data Mining in Pattern Recognition*, pages 86–95. Springer-Verlag, 2003. ISBN 3-540-40504-6. URL `http://dl.acm.org/citation.cfm?id=1759548.1759559`.

6. Maria Biryukov and Cailing Dong. Analysis of computer science communities based on dblp. In *Proceedings of the 14th European Conference on Research and Advanced Technology for Digital Libraries*. Springer-Verlag, 2010. ISBN 3-642-15463-8, 978-3-642-15463-8. URL `http://dl.acm.org/citation.cfm?id=1887759.1887792`.

7. T. Chakraborty, S. Kumar, M. D. Reddy, S. Kumar, N. Ganguly, and A. Mukherjee. Automatic classification and analysis of interdisciplinary fields in

computer sciences. In *2013 International Conference on Social Computing*, pages 180–187, Sep. 2013. doi: 10.1109/SocialCom.2013.34.

8. Tanmoy Chakraborty and Subrata Nandi. Universal trajectories of scientific success. *Knowl. Inf. Syst.*, 54(2):487–509, February 2018. ISSN 0219-1377. doi: 10.1007/s10115-017-1080-y. URL `https://doi.org/10.1007/s10115-017-1080-y`.

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

10. Emanuele Coviello, Antoni B. Chan, and Gert R. G. Lanckriet. Clustering hidden markov models with variational hem. *J. Mach. Learn. Res.*, 15:697–747, 2014. ISSN 1532-4435. URL `http://dl.acm.org/citation.cfm?id=2627435.2627457`.

11. Cristian Danescu-Niculescu-Mizil, Robert West, Dan Jurafsky, Jure Leskovec, and Christopher Potts. No country for old members: User lifecycle and linguistic change in online communities. In *Proceedings of the 22Nd International Conference on World Wide Web*. ACM, 2013. ISBN 978-1-4503-2035-1. doi: 10.1145/2488388.2488416. URL `http://doi.acm.org/10.1145/2488388.2488416`.

12. A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B*, 39:1–38, 1977.

13. Pierre Deville, Dashun Wang, Roberta Sinatra, Chaoming Song, Vincent D Blondel, and Albert-László Barabási. Career on the move: Geography, stratification, and scientific impact. *Scientific reports*, 4:4770, 2014.

14. Adabriand Furtado, Nazareno Andrade, Nigini Oliveira, and Francisco Brasileiro. Contributor profiles, their dynamics, and their importance in five q&a sites. CSCW'13, 2013. ISBN 978-1-4503-1331-5. doi: 10.1145/2441776.2441916. URL `http://doi.acm.org/10.1145/2441776.2441916`.

15. Shima Ghassempour, Federico Girosi, and Anthony Maeder. Clustering multivariate time series using hidden markov models. In *International journal of environmental research and public health*, 2014.

16. Cyril H Goulden et al. Methods of statistical analysis. *Methods of statistical analysis.*, 1949.

17. B.H. Juang and Lawrence R. Rabiner. A probabilistic distance measure for hidden markov models. 64, 02 1985.

18. Shulamit Kahn. Gender differences in academic career paths of economists. *The American Economic Review*, 83(2):52–56, 1993. ISSN 00028282. URL `http://www.jstor.org/stable/2117639`.

19. Bernhard Knab, Alexander Schliep, Barthel Steckemetz, and Bernd Wichern. *Model-Based Clustering With Hidden Markov Models and its Application to Financial Time-Series Data*. Springer Berlin Heidelberg, 2003. ISBN 978-3-642-18991-3. doi: 10.1007/978-3-642-18991-3_64. URL `https://doi.org/10.1007/978-3-642-18991-3_64`.

20. William H. Kruskal and W. Allen Wallis. Use of ranks in one-criterion variance analysis. volume 47, pages 583–621. Taylor & Francis, 1952.

21. S. Kullback and R. A. Leibler. On information and sufficiency. *Ann. Math. Statistics*, 22:79–86, 1951. ISSN 0003-4851. doi: 10.1214/aoms/1177729694. URL https://doi.org/10.1214/aoms/1177729694.

22. Himabindu Lakkaraju, Stephen H. Bach, and Jure Leskovec. Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 1675–1684. ACM, 2016. ISBN 978-1-4503-4232-2. doi: 10.1145/2939672.2939874. URL http://doi.acm.org/10.1145/2939672.2939874.

23. Lei Li and B. Aditya Prakash. Time series clustering: Complex is simpler! In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, USA, 2011. Omnipress. ISBN 978-1-4503-0619-5. URL http://dl.acm.org/citation.cfm?id=3104482.3104506.

24. Yong Liu, Jorge Goncalves, Denzil Ferreira, Bei Xiao, Simo Hosio, and Vassilis Kostakos. Chi 1994-2013: mapping two decades of intellectual progress through co-word analysis. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*, pages 3553–3562. ACM, 2014.

25. Helmut Ltkepohl. *New Introduction to Multiple Time Series Analysis*. Springer Publishing Company, 2007. ISBN 3540262393, 9783540262398.

26. Marcelo Maia, Jussara Almeida, and Virgilio Almeida. Identifying user behavior in online social networks. SocialNets'08, 2008. ISBN 978-1-60558-124-8. doi: 10.1145/1435497.1435498. URL http://doi.acm.org/10.1145/1435497.1435498.

27. Lena Mamykina, Bella Manoim, Manas Mittal, George Hripcsak, and Björn Hartmann. Design lessons from the fastest q&a site in the west. CHI'11, 2011. ISBN 978-1-4503-0228-9. doi: 10.1145/1978942.1979366. URL http://doi.acm.org/10.1145/1978942.1979366.

28. Julian John McAuley and Jure Leskovec. From amateurs to connoisseurs: Modeling the evolution of user expertise through online reviews. In *Proceedings of the 22Nd International Conference on World Wide Web*. ACM, 2013. ISBN 978-1-4503-2035-1. doi: 10.1145/2488388.2488466. URL http://doi.acm.org/10.1145/2488388.2488466.

29. Kanika Narang, Austin Chung, Hari Sundaram, and Snigdha Chaturvedi. Discovering archetypes to interpret evolution of individual behavior. *arXiv preprint arXiv:1902.05567*, 2019.

30. Lawrence R. Rabiner. Readings in speech recognition. chapter A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, pages 267–296. Morgan Kaufmann Publishers Inc., 1990. ISBN 1-55860-124-4. URL http://dl.acm.org/citation.cfm?id=108235.108253.

31. Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 1135–1144. ACM, 2016. ISBN 978-1-4503-4232-2. doi: 10.1145/2939672.2939778. URL http://doi.acm.org/10.1145/2939672.2939778.

32. Tara Safavi, Maryam Davoodi, and Danai Koutra. Career transitions and trajectories: A case study in computing. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, pages 675–684. ACM, 2018. ISBN 978-1-4503-5552-0. doi: 10.1145/3219819.3219863. URL http://doi.acm.org/10.1145/3219819.3219863.

33. Tiago Santos, Simon Walk, Roman Kern, Markus Strohmaier, and Denis Helic. Activity archetypes in question-and-answer (q8a) websites—a study of 50 stack exchange instances. *Trans. Soc. Comput.*, 2(1):4:1–4:23, February 2019. ISSN 2469-7818.

34. Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-June (Paul) Hsu, and Kuansan Wang. An overview of microsoft academic service (mas) and applications. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15 Companion, pages 243–246. ACM, 2015. ISBN 978-1-4503-3473-0. doi: 10.1145/2740908.2742839. URL http://doi.acm.org/10.1145/2740908.2742839.

35. Padhraic Smyth. Clustering sequences with hidden markov models. In *Advances in Neural Information Processing Systems*, pages 648–654. MIT Press, 1997.

36. Robert Tibshirani, Guenther Walther, and Trevor Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423, 2001.

37. Dashun Wang, Chaoming Song, and Albert-László Barabási. Quantifying long-term scientific impact. *Science*, 342(6154):127–132, 2013.

38. Melanie Ward. The gender salary gap in british academia. *Applied Economics*, 33 (13):1669–1681, 2001. doi: 10.1080/00036840010014445.

39. Samuel F. Way, Daniel B. Larremore, and Aaron Clauset. Gender, productivity, and prestige in computer science faculty hiring networks. In *Proceedings of the 25th International Conference on World Wide Web*, 2016. ISBN 978-1-4503-4143-1. doi: 10.1145/2872427.2883073. URL https://doi.org/10.1145/2872427.2883073.

40. Samuel F. Way, Allison C. Morgan, Aaron Clauset, and Daniel B. Larremore. The misleading narrative of the canonical faculty productivity trajectory. *Proceedings of the National Academy of Sciences*, 114(44):E9216–E9223, 2017. ISSN 0027-8424. doi: 10.1073/pnas.1702121114. URL http://www.pnas.org/content/114/44/E9216.

41. B. L. WELCH. The generalization of student's problem when several different population varlances are involved. *Biometrika*, 34(1-2):28–35, 01 1947. ISSN 0006-3444. doi: 10.1093/biomet/34.1-2.28. URL https://dx.doi.org/10.1093/biomet/34.1-2.28.

42. Jaewon Yang and Jure Leskovec. Patterns of temporal variation in online media. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, New York, NY, USA, 2011. ISBN 978-1-4503-0493-1. doi: 10.1145/1935826.1935863. URL http://doi.acm.org/10.1145/1935826.1935863.

43. Jaewon Yang, Julian McAuley, Jure Leskovec, Paea LePendu, and Nigam Shah. Finding progression stages in time-evolving event sequences. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW'14, pages 783–794. ACM, 2014. ISBN 978-1-4503-2744-2. doi: 10.1145/2566486.2568044. URL `http://doi.acm.org/10.1145/2566486.2568044`.