SYM-AM-22-042

Excerpt from the Proceedings

of the

Nineteenth Annual Acquisition Research Symposium

Acquisition Research: Creating Synergy for Informed Change

May 11–12, 2022

Published: May 2, 2022

Disclaimer: The views represented in this report are those of the author and do not reflect the official policy position of the Navy, the Department of Defense, or the federal government.

The research presented in this report was supported by the Acquisition Research Program at the Naval Postgraduate School.

To request defense acquisition research, to become a research sponsor, or to print additional copies of reports, please contact any of the staff listed on the Acquisition Research Program website (www.acquisitionresearch.net).

Predictability and Forecasting of Acquisition Careers in the Army

Eduardo López—primary expertise is in the application of complex network techniques to the study of real-world systems. This has been applied in a wide range of contexts that include patterns of human communication in social networks, search and matching processes on networks specialized to employment, biological nutrient transport in networks of fungi, and data traffic in data networks. From the methodological direction, López has worked on the basic properties of network transport, robustness, and percolation and temporal evolution, including single-layer as well as multiplex networks. López has been employed at the Physics Department at Universidad del Zulia as Faculty in Training (1996–2000), at the Physics Department at Boston University as a Research Assistant (2001–2005), at the Center for Nonlinear Studies at Los Alamos National Laboratory as a Postdoctoral Research Fellow (2005–2008), at the University of Oxford (2008–2012) as a Postdoctoral Fellow of Complexity Science and as a James Martin and Senior Researcher (2012–2016), and at the Department of Computational and Data Sciences as an Assistant Professor (2016–present). His degrees include a BA in physics (summa cum laude) from the Universidad del Zulia, Venezuela; an MA in physics from Boston University; and a PhD in physics from Boston University. [elopez22@gmu.edu]

Frank B. Webb—is a PhD student in George Mason University's Computational Social Science program working on modeling labor flow dynamics within organizations. Prior to beginning his PhD, his undergraduate degree was in Biological Sciences from the University of South Carolina where his research work focused on social media analysis for discussions of LGBTQ+ health on Twitter. This work was recognized with poster awards at the annual University of South Carolina research exhibitions. His research interests include systems modeling, networks in social systems, social media, and LGBTQ+ health care. [fwebb2@gmu.edu]

Abstract

A great deal is known about the movement of personnel population within large organizations (manpower). On the other hand, far less is known about how individual careers unfold through the structure of such organizations, with no established methods to forecast the positions individuals will take in so-called internal labor markets. In this paper, based on methods from network science, probability, and data analysis, we provide a new, empirically calibrated modeling framework for forecasting careers in large organizations. We show that, without the use of information that goes beyond the memoryless framework provided by Markov models, it is not possible to understand and forecast career moves in an organization. When memory effects are included, models improve significantly and begin to provide both useful predictions as well as information about the limits of predictability in career forecasting. Our method is applied to the Army acquisition workforce.

Introduction and Background

An effective way to summarize how organizations assign personnel to their tasks is offered by Bidwell (2017), who stated, "Perhaps the most basic challenge in talent management is ensuring that a company has the right people in the right places when it needs them." This succinct description contains a great deal of information. The places that Bidwell mentions represent positions in the organization responsible for certain tasks. Furthermore, the people that perform these tasks must possess the appropriate skills, training, experience, and social capital to be able to successfully complete these responsibilities. Seen from the lens of a mathematical description, both job positions and individuals are each represented by collections of attributes designed to capture the respective tasks, skills, experiences, and other important characteristics that can describe this system.

Mathematical descriptions of systems of this type have been studied in the past from the standpoint of the organization (De Feyter & Guerry, 2011; Wang, 2005) or the individuals embedded in it (see, for example, Stewman & Konda, 1983; Stewman & Yeh, 1991). This division reflects the multiscale perspective of the problem. The first literature, focused on *manpower* at the organizational scale, conceived the organization as made up of a set of fixed position types (similar to classifications or ranks) and personnel moves (also called stock moves) among these types. On the other hand, the individual perspective is connected to the literature in *career* studies (Gunz et al., 2020), sometimes also referred to as internal labor markets (Stewman, 1986), and is highly influenced by the idea that each employee inside the organization sporadically moves between vacancies that become available, effectively establishing two dynamic populations of vacancies and workers that interact with each other. The descriptions at either scale share many characteristics. First, they are predominantly stochastic in nature, almost always reliant on Markov models (or generalizations such as semi-Markov models; see, for example, Ginsberg, 1971). Second, they conceptualize the organization as static, which means that any temporal behavior is limited to the micro-dynamics of individual vacancies or individuals. Third, mostly due to lack of detailed micro-level information, they abstract much of the multidimensional information about the system such as the internal administrative structure of organizations (its subunits), the details of each job position, the social networks, the work teams, or other local behavior. The overall performance of these modeling approaches has been mixed: while the organizational level manpower literature has been able to offer rather reliable forecasts of personnel stocks in the system, the individual level literature has been less successful in predicting how people will move through the organization as they progress in their careers.

Much has changed over the past two and a half decades during which modeling questions took a back seat to other theoretical considerations that have occupied the research community studying careers in and out of organizations (for a discussion, see, for example, Bidwell, 2017). First, computational power and the availability of extensive data have transformed the way in which we view human-centric problems, where it is now feasible to consider modeling approaches that used to be found only in the physical sciences and engineering. Second, a new conceptual framework for highly complex and heterogeneous systems emerged in the form of the discipline of Complex Networks (see, for example, Barabási, 2014, for a popular presentation, and Newman, 2018, for a formal presentation), which provides a precise mathematical description of large interacting and heterogeneous systems such as human organizations. Both of these factors have played an important role in the development of a new theoretical view of job mobility, the concept of *Labor Flow Networks* (LFN), introduced for the purposes of modeling job changes with a simultaneous high-resolution and large system visibility (Guerrero & Axtell, 2013; Axtell et al., 2019; López et al., 2020).

An important consideration emerging from the LFN literature and other lines that have sprouted from it (see, for example, Mealy et al., 2018) refers to what is tracked in such career sequences. The traditional choice in career studies has been rank or some equivalent of it (see Rosenbaum, 1979; Stewman, 1986). The recent LFN literature focuses on firms (Guerrero & Axtell, 2013) due to their critical role in the economy and the fact that most approaches to the problem of job search have unfortunately ignored the firm scale. As will be shown, when enough information is available, there are multiple choices one has to track careers.

In this paper, by combining the LFN notion and stochastic processes with memory (Rosvall et al., 2014), we present a framework for tracking and modeling the movement of personnel through a large organization and apply the method to the Army acquisition workforce (AAW). The method seeks a clearer understanding of the formation of career sequences in an organization and how probable each sequence is. This information can be used to forecast future careers of interest to individual employees as well as the organization as a whole. We find that the introduction of memory dramatically increases the performance of a forecasting model, eliminating most of the unrealistic career sequences predicted by the current state of the art, while simultaneously generating better probability estimates of the number of employees actually choosing a sequence. Longer career sequences are generally less well predicted, although performance is still quite good. Our method also identifies career sequences that are unlikely to occur from the standpoint of what is known in the theory careers, and thus provides an opportunity to add new understanding to this field. Our results benefit greatly from the development of complex network techniques in recent decades.

We study two different definitions of career stops within the organization, operational units and occupational series (as defined by the U.S. Office of Personnel Management [OPM]). The study of the first type of stop (operational units) is an important addition that we bring to the literature, extending the notion of career sequences to the operating units/departments. This result follows a similar line of thinking as in the LFN literature. The relevance of this notion is that, while understanding a career in terms of the occupations says a great deal about skills, it says very little about social capital. On the other hand, career sequences tracked at operational units can carry social information in the form of personal contacts that are generated by directly working with others.

The first step in our framework involves an empirical analysis of personnel movements in the organization. This analysis yields a set of transition probabilities that can be used in either a memoryless form, the common approach in most personnel modeling, or by drawing on information about prior personnel movements in order to inform future ones. The second step in the framework involves a stochastic process that simulates how personnel would move inside the organization. To understand the impact of memory, our stochastic processes are chosen to be either first- or second-order Markov chains; first order chains are memoryless, whereas second order chains remember the most recent transitions before picking among subsequent choices. Although in some simple cases, these models could be solved mathematically, for almost any realistic data set, numerical approaches are needed to measure the statistics of outcomes. The third step in the framework corresponds to its evaluation, along with the tracking of any behaviour that deviates considerably from the predictions of the stochastic process. This evaluation is performed through the use of tools from information theory (Cover & Thomas, 2006) and statistics.

Our work creates a renewed opportunity to understand and forecast careers in organizations. In particular, by modifying the scale at which careers are studied, moving away from stocks of individuals progressing through ranks to looking at them in a more granular way, it makes it possible to bring into the picture other literatures such as that of quantitative career clustering, initiated by the work of Abbott and Hrycak (1990) and further perfected in subsequent decades (Aisenbrey & Fasang, 2010). Another line of this literature is the one initiated by Rosenbaum (1979), which provided the first empirical evidence from administrative records of history playing a role in the speed and attainment of career progression through a mechanism of tournaments (in a sense, highlighting the weakness inherent in memoryless models).

Materials and Methods

Data

The data we study is for the AAW and has two parts, one associated with individuals and the other with the structure of the AAW. The data sets cover the period between 2012 and 2020. All employee records are anonymized by associating to each individual a hashed key. Each employee record contains the position occupied on every month when the employee is part of the AAW. This information includes the operational unit of the individual as well as his/her occupational series (from the OPM classification). Over the period of the data, the AAW has ranged in size between under 35,000 to close to 42,000 individuals. There are around 1,000 operational units in the AAW, and employees span close to 100 occupational codes.

Methods

Our approach to career sequences in organizations deals with ordered chains. This literature emerged with White (1970), who realized their role in careers, paying special emphasis to the notion of vacancy chains. A vacancy chain emerges when a person leaves the post they are occupying to take a new job inside or outside an organization, leading to another person eventually occupying the vacancy but creating a new one, and so on. The successive vacancies created are called vacancy chains. Subsequently, a broader notion of social sequences has emerged that spans well beyond careers (see Cornwell, 2015) and has gained traction in the mathematical sociology literature.

The quantities we use in this paper are formal, and their detailed definitions are given in the Appendix. Here we merely introduce the notation and explain the spirit of these quantities. The application of these quantities in evaluating our models is done in the Results section.

As mentioned in the Introduction, career modeling is based on stochastic processes. Our approach is to use the data from observed job transitions to create information about future transitions, specifically, probabilities for such transitions. The spirit behind this idea is supported by the work of Collet and Hedström (2013) and López et al. (2020), which shows that once a job transition is observed between two firms in an open economy, the chance that any new random occurs between those two firms is about 1,000 times larger than between two firms without any previous transitions. This notion led Guerrero and Axtell (2013) to define the LFN.

Therefore, to capture the probabilities of transitions an individual may have of performing a particular job change in the AAW, we define two versions of transition probabilities, one for the model that ignores memory, and another for the model with memory of the most recent job change. These two quantities are, respectively, $p_{l,g}$ and $p_{(l,g),(g,h)}$, where *l, g,* and *h* all represent stopping points along a career sequence. Note that such transition probabilities are the result of an aggregation of the actions by many people going through job changes in the AAW over a period of time. Hence, this captures a notion of popular moves.

Note that, as in the LFN literature, we think of an organization as structured into such stops, connected if there have been job transitions between those stops. The stops can represent, for example, the occupational series a person has while in a job, and the career sequence is a sequence of occupations. Another stop could be the operational unit to which the employee is attached while having a post, and in this case the career sequence is a sequence of operational units. The stops can be other concepts as well. Critically, those stops are equivalent to nodes in a network, while the connections between the nodes represent observed job changes.

The models we apply make use of the transition probabilities stated above when an employee decides to change jobs, either recalling or ignoring the previous job change it performed.

Another quantity we rely on is $\square_l$, which are the chances that somebody at stop/node *l* separates (decides to leave or is told to do so) from their current position in any given interval of time. Individuals will not have to decide on where to go unless they separate from their current position. In our model, each time interval is of 1 month.

Models then generate *in silico* careers (simulated in the computer). One can generate as many as desired and, in fact, this is needed since job changes all have an element of randomness. With careers starting at a node *l*, we track all the career sequences *observed* from that starting node; some are performed by multiple people, some by just one. This allows us to create probabilities $F(S|s_1=l)$ of observing a specific career sequence S that started at stop $s_1 = l$. In a similar way, the careers we simulate have probabilities captured in $\square(S|s_1=l)$. Both $F(S|s_1=l)$ and $\square(S|s_1=l)$ are examples of so-called probability distributions. The quality of a model is assessed by the similarity that $F(S|s_1=l)$ and $\square(S|s_1=l)$ may have.

Because these probabilities have multiple parameters, we check for their similarity in multiple ways. We check if they produce exactly the same sequences. This is done by a quantity called the Jaccard index. We also check if the probabilities assigned to the careers that are both simulated and observed are similar, independent of whether all observed careers are generated in our models. This is done in two ways. One is based on a concept from information theory called the Jensen–Shannon Divergence (*JSD*; Lin, 1991), which measures the number of bits (in terms of information in a computer) that separate the two probability distributions (observed and simulated). The other is based on a comparison of career sequence by career sequence, that is, $\square(S|s_1=l)/F(S|s_1=l)$ for every S in starting from every l. The closer these ratios are to 1, the better the model. To assess this proximity to 1, we introduce a final set of variables, the most important of which is called Var($\square_l$), capturing the cumulative deviations from one of the logarithms of the ratio of probabilities. Basically, the bigger this number is, the worse the model is doing.

As an important technical point, the use of random simulations means that we do not typically generate the same $\square(S|s_1=l)$ in every simulation. This means that comparisons between $\square(S|s_1=l)$ and $F(S|s_1=l)$ are actually done between the latter and a whole set of samples of the former (for which we use a labelling index r). In particular, we calculate *JSD* between pairs of distinct simulations of $\square(S|s_1=l)$, with each result being labelled $t_{r,r'}$. We also calculate *JSD* between $\square(S|s_1=l)$ and $F(S|s_1=l)$, with each result being labelled m_r.

With or without memory, this network construction based on previous job transitions does a good job of modeling the career sequences of the system, although memory makes the results considerably better in some key ways. On the other hand, the probability ratios allow us to spot career sequences that are inherently difficult to model, which we briefly discuss.

Results

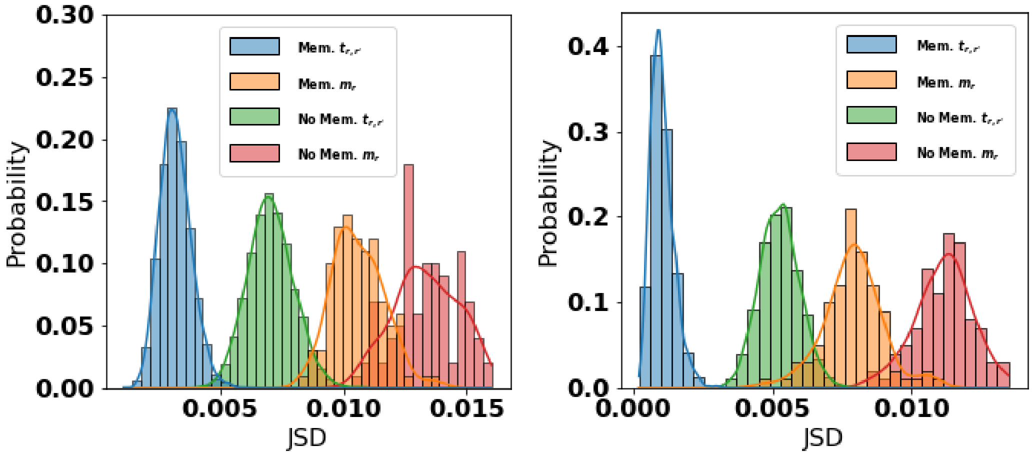
We divide the presentation of our results into two career sequences defined on occupation nodes or on operational units. As the results show, there is great consistency between the two.

Occupational Model Results

We first focus on the application of the model to occupational series. In this case, each stop corresponds to the occupational series an employee has upon being first observed in the data. The working unit of this employee is not considered in this analysis.

In Figure 1, we present results for the *JSD* distributions ($m_r(l)$ and $t_{r,r'}(l)$) for two different starting occupations l. The two occupations are 0346 (Logistics Management Series) and 0802 (Engineering Technical Series). The values of *JSD* generated from the models are considerably small, indicating their general quality. Moreover, comparing the memoryless and the one-step memory models, in both examples we see how the latter model performs better than the former. This is not a feature of these two occupations as comprehensive exploration of all occupational series leads to the same result.

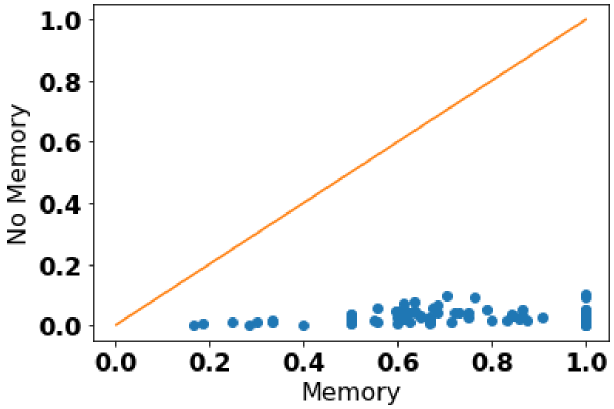
An interesting effect to explain is the fact that the values $m_r(l)$ are typically larger than the values $t_{r,r'}(l)$. This is due to the fact that while the samples $\square_r(S|s_1=l)$ from any of the models are self-consistent (one value of r is similar to another one r'), the consistency between each $\square_r(S|s_1=l)$ and $F(S|s_1=l)$ is generally less. In other words, the models are good but not perfect. In our analysis, we do find some occupations where the simulations match the data well, but this is not guaranteed.



Note. Each panel is associated with an initial occupation (in this case 0346 on the left and 0802 on the right). In each panel, four distributions are displayed, differentiated by color (blue corresponds to $t_{r,r'}(I)$ for the model with memory, green corresponds to $t_{r,r'}(I)$ with no memory, orange corresponds to $m_r(I)$ with memory, and red corresponds to $m_r(I)$ with no memory). In all cases, the distribution of $m_r(I)$ peaks at smaller values (of *JSD*) for the model with memory.

Figure 1. Probability Distributions of $m_r(I)$ and $t_{r,r'}(I)$ for Two Different I, and Models

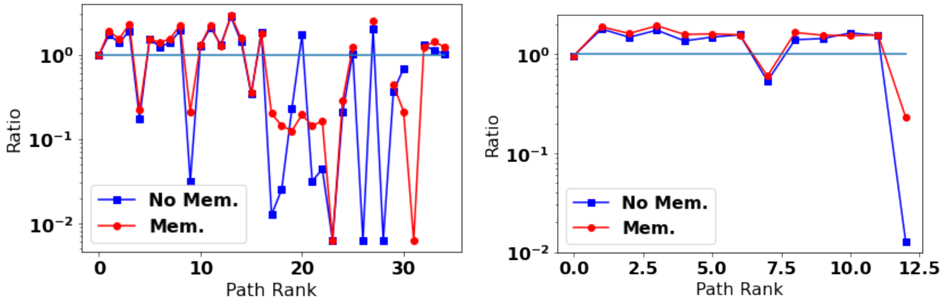
As explained in the Methods section, *JSD* captures an aggregate measurement of the discrepancy between $F(S|s_1=I)$ and the models. However, other differences between model and $F(S|s_1=I)$ can remain unseen in this analysis. The most critical of those features is the possibility that a model generates career sequences that do not always reflect well the collection of observed careers. These possible differences can be assessed by the Jaccard index (see Figure 2). The results of this analysis clearly show the considerable improvement brought on by the introduction of memory: while the values of the index for the memoryless cases remain bounded from above by a value near 0.1, the one-step memory leads to indices with values ranging from about 0.2 to 1.0. If there were a strong correlation between the Jaccard indices of the models for the same starting occupations, the points would lie near the reference line along the diagonal, but this is not the case. The main reason why the Jaccard index improves so dramatically is because the number of careers generated in the one-step memory model is considerably smaller than in the memoryless model. Furthermore, the generated careers generally capture the observed careers, making it possible for the index to reach values that tend to 1.



Note. Each point corresponds to a starting occupation and has as its horizontal coordinate the value of the Jaccard index for the one-step memory model, and as its vertical coordinate the value of the Jaccard index for the memoryless model. The modeled careers correspond to the union of all careers created in the n_w $n_\square$ total number of walks across all $n_\square$ realizations. While the values of Jaccard indices for the one-step memory span the range between approximately 0.2 to 1.0, the Jaccard indices of the memoryless model remain quite low, usually no larger than 0.1. The orange line runs along the diagonal as a visual reference. If the two models provided similar values of Jaccard indices, one would expect to see the cloud of points near that line.

Figure 2. Scatter Plot of the Jaccard Indices Between the Memoryless and One-Step Memory Models, Calculated Between the Careers Generated From Simulations and From Observation

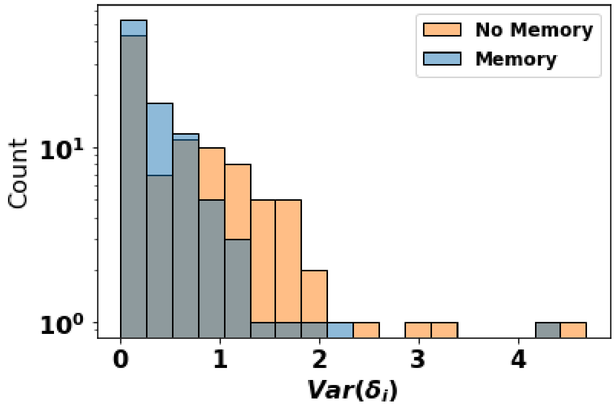
Both *JSD* and Jaccard indices produce a summary statistic about the details of the relationship between $F(S|s_1=l)$ and the model outputs captured in the realizations $\square_r(S|s_1=l)$. As defined in the Methods section, a more direct analysis of each career sequence S that belongs to these distributions can be achieved through $d(S_i)$. For a given l, we plot $(i, d(S_i))$ for the two models. This is shown in Figure 3. The blue and red curves present, respectively, the memoryless and one-step memory models. Generally, for any career S_i, $d(S_i)$ is closer to 1 for the one-step memory model, which is desired.



Note. The blue curve represents the memoryless model, whilst the red curve represents the one-step memory model. Both plots also show a horizontal line of height 10^0 (i.e., 1), which is the target achieved for an "ideal" model that reproduces careers perfectly. The red curve is generally closer to the horizontal line = 1 for both starting occupations.

Figure 3. Profiles of Models in Terms of Their Relation to Observed Careers, Captured in $(i,d(S_i))$, for Starting Occupations 0346 and 0802

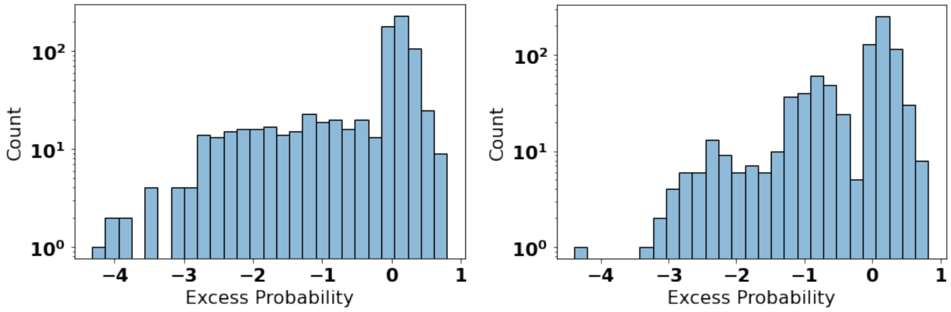
The results offered by the analysis of $d(S_i)$ are limited in that they require l by l analysis. However, it is desirable to quantify all l in a systematic way, which was the reason for the introduction of $\square_l$ and its variance Var($\square_l$). This last quantity can be studied for the entire system through its histogram (see Figure 4).



Note. The histogram clearly shows the larger values corresponding to the memoryless model.

Figure 4. Histogram of Var(□$_i$) for Both the Memoryless and One-Step Memory Models for Occupations

A final analysis comes from a careful study of Figure 3, which illustrates examples of careers that, while simulated by the random models, appear with frequencies far different than observed. This is reflected in the values of $d(S_i)$, which, due to the clarity of its interpretation, we analyse as $|\log d(S_i)|$. When one of these values exceeds some arbitrarily chosen threshold, career sequence S_i is taken to be significantly outside the model. To first develop a notion of the possible values that $\log d(S_i)$ can take, we present Figure 5 for the two models and across all careers in the system. It is clear that the majority of the career sequences have values of log $d(S_i)$ in the vicinity of 0 and < 1. On the other hand, both models have a relatively long tail of values below 0, which means particular career sequences observed in the data are not simulated as often in the models. The memoryless model shows even more careers that significantly deviate from their observed frequencies than the model with memory. In addition, the one-step memory model shows multi-modality (although we do not present this analysis, we have traced this result to career sequence length, i.e., longer careers are harder to model accurately).



Note. The model with memory concentrates more of its probability mass between −1 and 1 than the memoryless model.

Figure 5. Histogram of Values of log $d(S_i)$ for the Memoryless and One-Step Memory Models for Occupational Career Sequences

On the basis of the results in Figure 5, we see that considerable interesting behaviour occurs when $\log d(S_i) \leq -1$. Are there common features to careers that cross this threshold? One feature, mentioned above, involves the length of paths. Longer career sequences are harder to forecast and lead to values of □$(S|s_1=l)$ that deviate from $F(S|s_1=l)$ more. Beyond this length effect, other details of career sequences may be responsible for leading to poor forecasts from models.

A comprehensive exploration of career sequences in order to identify all possible reasons behind poor predictions of those sequences is not likely to be very informative, as any single sequence can have its own reasons for being hard to predict. A more productive approach may be to identify temporal features shared by many poorly forecasted careers so that especial approaches can be applied to improve those forecasts. The concept of a temporal pattern in a network is known as a temporal motif (Holme & Saramäki), and our method for understanding poorly forecasted careers is basically a search for such temporal motifs.

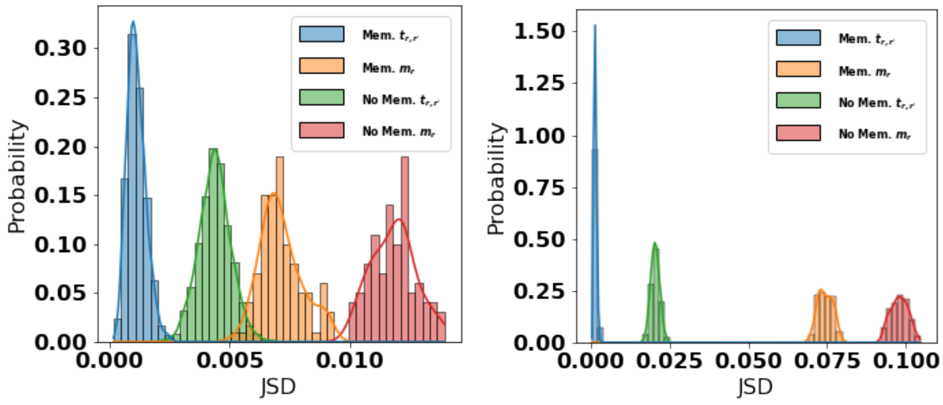
Our analysis has yielded an interesting and unexpected result. One of the key temporal motifs contributing to poor prediction is one characterized by employees going back to positions they previously held. This is a surprising result. In the observed paths, 24.5% contain this motif. Without memory, the model was able to produce 79.6% of those motif paths while memory improved this to 88.2%.

In summary, the results for the analysis of careers sequences defined on occupations shows that the models we have constructed are certainly useful and, furthermore, that the one-step memory model performs better.

Units Model Results

The approach deployed for the study of occupations can also be applied to the study of careers occurring along operational units of the organization. Methodologically speaking, there is no difference in the calculation of the quantities presented above, but interpretation of the results has to take into account the nature of the nodes. Qualitatively speaking, we find the same behavior in career sequences tracked on the basis of operational units as we observe for sequences over occupational series.

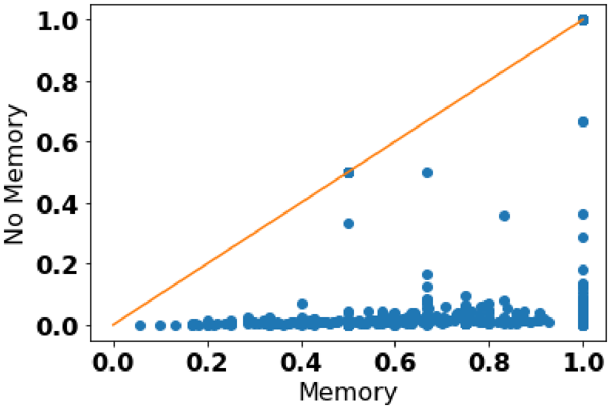
As an illustration of the similarity between modeling by occupational series or operational units, we present Figure 6, which shows the *JSD* measurements of careers starting from two such units. The observed features of these plots do not differ from those in Figure 1, that is, better performance for the one-step memory, as well as the observation that both models still have room for improvement.



Note. Each panel is associated with an operational unit (kept undisclosed). In each panel, four distributions are displayed, differentiated by color (blue corresponds to $t_{r,r'}(l)$ for the model with memory, green corresponds to $t_{r,r'}(l)$ with no memory, orange corresponds to $m_r(l)$ with memory, and red corresponds to $m_r(l)$ with no memory). In all cases, the distribution of $m_r(l)$ peaks at smaller values (of *JSD*) for the model with memory.

Figure 5. Probability Distributions of $m_r(l)$ and $t_{r,r'}(l)$ for Two Different l, and Models

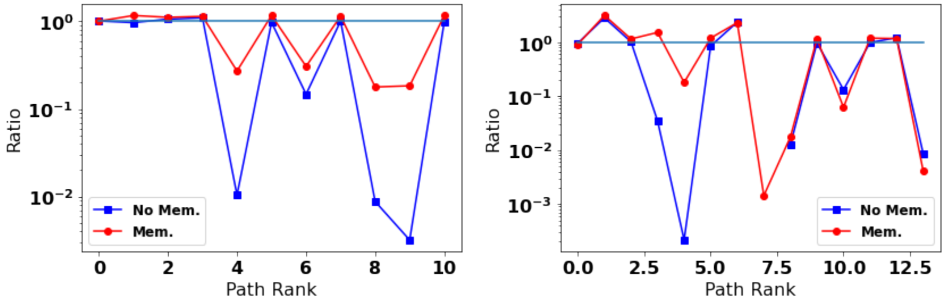
Next, we discuss the Jaccard over operational units. In contrast to the case of occupational series, there are a very distinct few units for which the memoryless model leads to good Jaccard indices (Figure 7). However, this is the exception rather than the rule. Overwhelmingly, the one-step memory model performs much better.



Note. Each point corresponds to a starting operational unit and has as its horizontal coordinate the value of the Jaccard index for the one-step memory model, and as its vertical coordinate the value of the Jaccard index for the memoryless model. The modeled careers correspond to the union of all careers created in the n_w $n_{\square}$ total number of walks across all $n_{\square}$ realizations. While the values of Jaccard indices for the one-step memory span the range between approximately 0.2 to 1.0, the Jaccard indices of the memoryless model remain quite low, usually no larger than 0.1. The orange line runs along the diagonal as a visual reference. If the two models provided similar values of Jaccard indices, one would expect to see the cloud of points near that line.

Figure 7. Scatter Plot of the Jaccard Indices Between the Memoryless and One-Step Memory Models, Calculated Between the Careers Generated from Simulations and from Observation

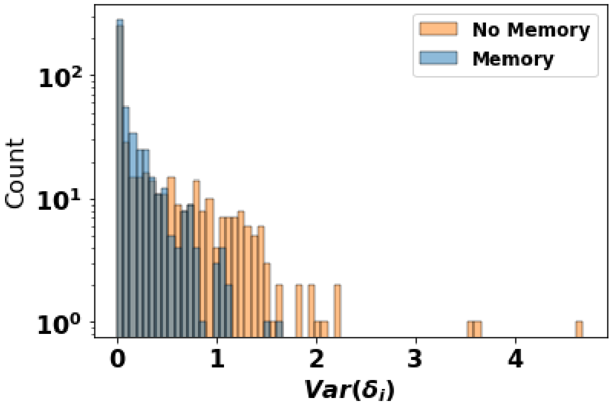
Results connected to $d(S_i)$, Var($\square_l$), and log $d(S_i)$ also have the same qualitative features for units as they do for occupations. For $d(S_i)$, we present Figure 8, which is constructed with the same units as in Figure 6. As for occupations, the match of the one-step memory model and observation is quite reasonable.



Note. The blue curve represents the memoryless model, whilst the red curve represents the one-step memory model. Both plots also show a horizontal line of height 10^0 (i.e., 1), which is the target achieved for an “ideal” model that reproduces careers perfectly. The red curve is generally closer to the horizontal line = 1 for both starting occupations.

Figure 8. Profiles of Models in Terms of Their Relation to Observed Careers, Captured in $(i, d(S_i))$, for Two Starting Undisclosed Operational Units

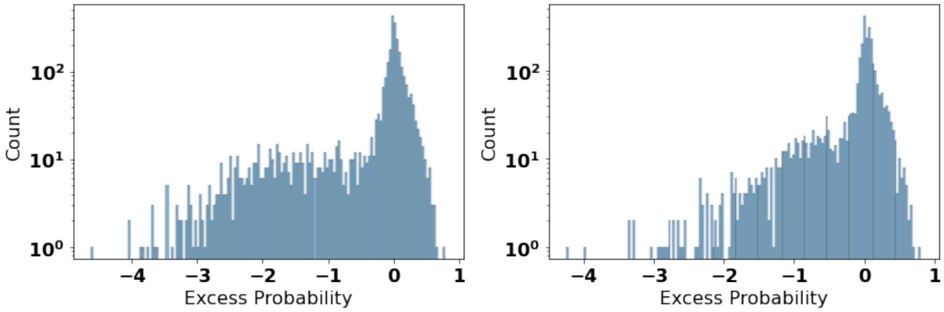
The values of Var($\square_i$) are shown in Figure 9.



Note. The histogram clearly shows the larger values corresponding to the memoryless model.

Figure 9. Histogram of Var($\square_i$) for Both the Memoryless and One-Step Memory Models for Operational Units

Finally, we present results for the collection of all log $d(S_i)$ in Figure 10.



Note. The model with memory concentrates more of its probability mass between −1 and 1 than the memoryless model.

Figure 10. Histogram of Values of log $d(S_i)$ for the Memoryless and One-Step Memory Models for Operational Units Career Sequences

As in the occupational series model, the motif of employees leaving one state for another and then returning to it is present in the units implementation. Here, 10.8% of the observed paths exhibit this motif. The model without memory is able to produce 88.2% of paths containing the motif. However, with memory, the model captures 98.1% of such paths.

Discussion and Conclusions

The modeling approach we have taken in this work has been highly driven by statistical analysis. The network structure implicit in the transition probabilities specified above (either for memoryless or one-step memory models) creates a network substrate that allows us to generate forecasts of the workforce job changes at a microscopic level, that is, for any career sequence.

The introduction of a notion of career sequences occurring on a network of operational units is new in the study of careers, and we expect that as we focus more on its details, numerous relevant features of the system will start to emerge, such as the value of work or friendship ties in people's careers.

An important limitation of our current methodology is that it is calibrated against observed job transitions rather than *possible* job transitions. This is an important issue because the finite nature of the system does not make it possible to observe enough job transitions that a probability for *any* arbitrarily chosen pair of transitions to occur can be extracted from the data. In order to overcome this, study of the characteristics of each job (say, occupational series, location, career field) offers a new direction to pursue in order to create a more flexible model that may be able to predict what could happen even if it has never been observed.

From the standpoint of the contribution that this work may bring to the acquisition workforce, we note that the Department of Defense requires the ability to understand high volumes of behavioral and environmental data, in an institutionally informed framework, to produce reliable forecasts of workforce behaviors across an extended planning horizon. This goal is consistent with the fact that one of the three priorities in the 2019 National Defense Strategy is to reform the department's business practices for performance and affordability (Mattis, 2018). Further, *The Army People Strategy* calls for the implementation of 21st century talent management, enabled by leading-edge research and leveraging technology and "data-driven organizational research to continuously improve Army people programs and policies" (Secretary of the Army, 2019).

Improving understanding about the way the government workforce moves within and across different organizations in detail, how to plan for it, and how to optimally manage it are clearly relevant strategic resource usage and institutional effectiveness concerns. Mission completion across the board is impacted directly by government organizations' ability to ensure capable people are in the right place, at the right time to perform critical tasks. In addition, findings from our ongoing research offer the promise of expanding the body of knowledge and theory of processes, systems, and policies both inside and outside the government.

In conclusion, our method allows us to create reliable forecasts of career sequences, especially as the memory of the model is increased. We expect this method to become useful in the near future as a forecasting tool for career moves inside the AAW. Longer term, we expect to develop a more extensive characterization of the forecasting power and limitations of this model.

Appendix: Formal Definitions

In order to provide some concrete definitions and notation, let us consider one hypothetical career sequence $S = (s_1, s_2, \ldots s_n)$. Each transition between two stops s_i and s_{i+1} provides information that the second of these stops can be reached by individuals in the first. Each stop s_i for all i of any career sequence takes its value from a set of allowed stops L, where the elements of L correspond to the kind of stop we are interested in modeling. For example, if we want to model movement of individuals through the units/departments of the organization, the elements of L will be the distinct organizational units; if we care about individuals moving through occupations, the elements of L will be occupational series codes and so forth.

To model careers, we define a stochastic process following the logic in López et al. (2020). An individual currently located in stop l has a decision to make: either remain in l with probability $1 - \square_l$, or depart with probability $\square_l$, where $l \in L$. Each element of an individual career s_i corresponds to a stop such as l. Another possible action that an individual can take is to exit the organization. In our approach, this is predetermined at the outset of an individual's career by assigning it a total time in the organization. Once the time assigned to the individual has elapsed, or the stipulated duration of the model is reached, the individual disappears.

Transitions between stops occur with some probability. In order to calibrate our model, we use information from observed careers. Using Q to denote the total number of sequences observed in our data, we can create a set of transition probabilities for our model by counting the number of individuals performing a given transition. As explained in the Introduction, we employ two different rules. First, in the case where prior transitions by an individual are considered irrelevant, we use the transition probability

$p_{l,g} = n_{l,g}/\square_g\, n_{l,g}$ [memoryless case],

where $l,g \in L$. In other words, the likelihood that an individual currently in l will transition to g as its next stop is given by the proportion of individuals in the past that, upon leaving l, decide to move to g.

The second type of transition probability we employ keeps track of the last transition made by an individual (if the individual indeed has a career sequence spanning at least one transition up to that point). In this case, we define the transition probability as

$p_{(l,g),(g,h)} = n_{(l,g),(g,h)}/\square_{(g,h)}\, n_{(l,g),(g,h)}$ [one-step memory case],

where $l,g,h \in L$ and (l,g) and (g,h) are transitions. The denominator sums over all possible destinations $h \in L$ that an individual that has arrived at stop g from stop l has been seen to reach. This case allows for the possibility that $g = h$, that is, that g is a terminal node for an individual that has reached g from l.

Two other rules apply to the model with memory. An individual for which l is their first stop*,* if they decides to change jobs, they does so under the rules of the memoryless model on this first change. This is because at that point, such individual does not have any prior history in the system to draw from. The second rule, already hinted at in the previous paragraph, is that memory can lead an individual to remain in a location due to their history. This is the case in which $p_{(l,g),(g,h)} = 1$ when $g = h$, because it means that in the data all those that arrived at g from l never moved away from g*.*

In both cases above, the process is Markovian in nature, as they abandon the memory of more remote events in the past. Extending memory is, in principle, straightforward, although computationally costly. However, as we shall see, single memory is sufficient to provide a strong predictive value to the model.

Both the memory and one-step memory processes described above can be encoded in a complex network, that is, an object in which every stop $l, g, h, \ldots$ can be thought of as a node of the network, and every transition between two stops (nodes) can be considered a link between the nodes. Thus, both $p_{l,g}$ and $p_{(l,g),(g,h)}$ lead to sets of nodes and links that represent the entire organization and its job transitions in the form of a network. When the nodes correspond to occupational series, the network is one of occupations and transitions between those occupations; when the nodes correspond to operational units, the network represents those operational units and the job transitions that occur across them. Given this interpretation of the model, in what follows, we interchangeably use nodes or stops to refer to either an occupation or unit of the organization.

In order to evaluate our models, we must compare their behavior to that of observed career sequences. Since the entry point of a career may play a role in its subsequent progression, we define a probability distribution $F(S|s_1{=}l)$ for all *observed* career sequences that share the same initial stop l. Thus, $F(S|s_1{=}l)$ is the probability that an individual that begins a career at l indeed performs the career sequence S. Our models also generate career sequences with some probability. We denote the probability distribution of *simulated* career sequences by $\square(S|s_1{=}l)$. Note that, in contrast to $F(S|s_1{=}l)$, $\square(S|s_1{=}l)$ is not fixed in our model.

This is because every time we construct a set of paths through a random process using a stochastic (Monte Carlo) computer simulation, the specific set of sequences and their relative proportions can be different. By the Law of Large Numbers, the larger the simulation in terms of the number of samples created, the less difference one expects between two separate Monte Carlo simulations, but it is very unlikely that for even a moderately large system one will obtain the same □($S|s_1=l$) twice. The detailed way in which examples (also called realizations) of □($S|s_1=l$) are constructed is explained in the Results section.

A model that is both perfect and can be simulated an infinite number of times would lead to $F(S|s_1=l)$ = □($S|s_1=l$) for any starting l, where both probability distributions would be defined over the same set of sequences. However, no model is perfect nor can it be simulated an infinite number of times. In order to measure the discrepancy between F and □, we employ three complementary methods. The first of these tracks how different the sets of sequences from each of the models are in comparison to the actually observed sequences. For this we introduce the notation $G_l=\{S|s_1=l, S$ observed$\}$ to represent the set of all observed career sequences that being at l (i.e., where the first stop s_1 is l). Similarly, we use $H_l=\{S|s_1=l, S$ simulated$\}$ to represent the set of simulated sequences. Then, we define the so-called Jaccard index

$$J = |\, G_l \cap H_l \,| \,/\, |\, G_l \cup H_l \,|,$$

where $G_l \cap H_l$ corresponds to the set intersection of G_l and H_l, and $G_l \cup H_l$ represents their union. Furthermore, the symbol | | measures the number of elements of a set. Thus, J measures the ratio of the number of common elements between G_l and H_l versus the total number of distinct elements contained in G_l and H_l. If $G_l=H_l$, $J = 1$, and if the two sets have no common elements, $J = 0$. Therefore, with the Jaccard index, we seek to determine if a model produces similar career paths, regardless of the rate (i.e., probability) at which they may be produced.

The second measure we employ in evaluating our models is the *JSD* (Lin, 1991), based on ideas from information theory. Whereas the Jaccard index captures the unweighted similarity between collections of sequences, the *JSD* measures "distance" between distributions. The units of *JSD* are basically those of information (i.e., bits). To interpret *JSD* results, it is useful to recall that a bit measures the information needed to describe something; more bits means more information needed. Now, since *JSD* is a distance between distributions, one expects that two identical distributions would have a *JSD* with a value of 0; distributions that are not equal will have a *JSD* > 0. The concrete definition of *JSD* requires the use of the concept of information entropy, that is, Shannon entropy *H,* which measures the number of bits needed to describe a probability distribution. Symbolically, it is given by Cover and Thomas (2006)

$$H(P) = -\ □_r\, P(r) \log_2 P(r),$$

where $P(r)$ is a probability distribution of some random variable r. A large value of $H(P)$ means that the distribution $P(r)$ requires a large amount of information to be described.

For the definition of H, we can now introduce the *JSD* we use. In particular, the *JSD* between $F(S|s_1=l)$ and an example of □($S|s_1=l$) is defined as

$$JSD(F,□) = H[(F+□□□□] – [H(F)+H(□)]/2.$$

As explained above, *JSD* acts as a distance in bits between probability distributions. In this specific case, the distance is measured between each distribution and the average distribution (F+□□□□. Ultimately, the intuition of how the value of *JSD* changes is clear: the more the difference between F and □□□the larger *JSD* becomes. Its lower bound is 0, but there is no upper bound in principle, although for any finite system, an upper bound could be found.

Note that because *JSD* is based on entropy, which, in turn, is calculated from probability distributions, the relative differences in likelihoods of career sequences are captured in this measure. However, because entropy is a sum, it does not keep track of which career sequences are the ones responsible for the most important contributions to *H* or *JSD.* For this reason, we need another measure.

In order to simultaneously address differences in probabilities between observed and simulated career sequences one sequence at a time, we introduce a graphical method that allows us to study discrepancies between $F(S|s_1=l)$ and examples of $\square(S|s_1=l)$. However, before we can deploy this approach, we require a prior step.

The career sequences generated by our models are not always the same as those observed. This occurs due to model stochasticity. Furthermore, the less accurate the model is, the more likely it is that observed and simulated careers differ. For the method we will present next, which focuses on the comparison of observed and simulated probabilities of career sequences, it is useful to correct $\square(S|s_1=l)$ so that it is conditioned on only those careers that are also observed. Therefore, we define

$$\square(S|s_1=l, S \text{ observed}) = \square(S|s_1=l) \,/\, \square_{S' \text{ observed}}\ \square(S'|s_1=l).$$

This expression creates a conditional probability of the simulated careers that are also observed careers.

We are now ready for the next analytical approach, which we first apply as a graphical method and then define from it a quantity that tracks difference so that we can systematically evaluate each modeling method over the entire set of sequences departing from any *l*. To be concrete, we define an ordered set of values $d(S_1),d(S_2),\ldots,d(S_{bl})$ where each S_i for *i* between 1 and b_l is taken to be an *observed* career sequence, and b_l is the number of them that have been observed starting at *l.* Then, any $d(S_i)$ is defined by

$$d(S_i) = \square(S_i|s_1=l) \,/\, F(S_i|s_1=l)\ [F(S_1|s_1=l) \geq F(S_2|s_1=l)\ )\ldots \geq F(S_{bl}|s_1=l)],$$

where the square brackets stipulate that the career sequences are indexed with *i* so that the sequence with the largest probability to be observed in the data is S_1, the sequence with the second largest probability to be observed is S_2, and so forth. Note that, if $\square(S_i|s_1=l) = F(S_i|s_1=l)$, for career S_i then $\square(S_i|s_1=l) / F(S_i|s_1=l) = 1$. Thus, we will be trying to evaluate the quality of each model by measuring how close to 1 the values $d(S_i)$ are.

The collection of $d(S_i)$ have another use: they are helpful in determining career sequences that substantially deviate from their observed probabilities. Thus, as part of our analysis, we track sequences that exceed some arbitrarily chosen threshold of deviation (specifically, we do this through $\log[d(S_i)]$, as explained later).

A given ordered set $d(S_1)$, $d(S_2)$, …, $d(S_{bl})$ can be plotted as a set of points $(i, d(S_i))$. This produces for each starting location *l* a *profile plot* that indicates how well each of the career sequences out of *l* have been captured by a model. This is a visual method to evaluate the models, one *l* at a time. However, to explore the entire set of all possible *l,* we cannot rely on visual inspection always, especially if careers are being studied using stops that are quite numerous in an organization (e.g., each job post). To address this, we introduce

$$\square_l \square\square\square \square_i \log[d(S_i)] \,/\, b_l\,,$$

$$\mathrm{Var}(\square_l) = \square_i\, (\log[d(S_i)] - \square_l)^2 \,/\, b_l\,,$$

which compute a measure of deviation between $F(S|s_1=l)$ and examples of $\square(S|s_1=l)$ combining the effect of all S_i. The use of the logarithm has a nice property from the standpoint of interpretation: when $\square(S_i|s_1=l) = F(S_i|s_1=l)$, $\log[\square(S_i|s_1=l) / F(S_i|s_1=l)] = \log(1) = 0$. The second of the two quantities, Var($\square_l$), captures a deviation from 0, and can only be positive, whereas $\square_l$ may have cancelations included (due to positive and negative values of $d(S_i)$). Therefore, we see Var($\square_l$) as a more robust quantification of deviation.

As indicated in the Methods section, each $\square(S|s_1{=}l)$ is generated through computer simulation in which in silico employees remain at a stop s_i with a probability $1 - \square_{si}$, move to another stop based on a probability $\square_{si}$ multiplied by a transition rate p that may depend on s_i only in the memoryless case or in the latest transition (s_{i-1}, s_i) in the one-step memory, and will exit the network after a number of predetermined steps □□□the latter being drawn from a distribution for the entire workforce. These rules will not always produce the exact same career sequences emerging from an initial stop l. Therefore, in order to generate a realization of $\square(S|s_1{=}l)$ that is representative of transitions observed, we create a number n_w of career walks. Usually, we employ 5,000 such career walks for a single $\square(S|s_1{=}l)$.

This, however, is not entirely sufficient to perform our analysis. If we consider the process of determining *JSD*, a single example of $\square(S|s_1{=}l)$ produces a single value of *JSD.* To provide us with enough statistics to understand the possible values of *JSD*, we generate $n_\square$ such simulations. We label the different realizations of $\square(S|s_1{=}l)$ emerging from this approach through the index r, that is, $\square_r(S|s_1{=}l)$, where r =1, …, $n_\square$. We then compute *JSD* in two different ways. To determine the similarity between model outputs and $F(S|s_1{=}l)$, we create $n_\square$ samples of *JSD* and label them $m_r(l)$, each given by

$$m_r(l) = JSD[F(S|s_1{=}l),\square\square_r(S|s_1{=}l)].$$

This generates a histogram of $n_\square$ values, each providing a *JSD* value between observed and simulated career sequences. The histograms are shown below in the subsections concerned with whether stops are defined as occupational codes or operational units.

Our second use of the $n_{\square\square}$realizations is to create a notion baseline value of *JSD* between simulations. This is done through the variable

$$t_{r,r'}(l) = JSD[\square_r(S|s_1{=}l)],\square\square_{r'}(S|s_1{=}l)]\ [r,\ r' = 1,\dots,\ n_\square,\ r \neq r'].$$

These pairwise combinations of the outputs of simulations r and r' lead to a total of $n_\square$ ($n_\square$ - 1)/2 values $t_{r,r'}$. We study these values using probability distributions as well (see Figure 1 and Figure 6). Usually, we employ $n_\square$ = 100, with $n_\square$ ($n_\square$ − 1)/2 = 100 × 99/2 = 4,950. This generates per l a total of 100 samples of $m_r(l)$ and 4,950 samples of $t_{r,r'}(l)$.

Why do we need to create samples for $m_r(l)$ and $t_{r,r'}(l)$? The overall reason is that they create ways of comparing the performance of the models against each other, and also as a way to determine if any of these models is actually achieving the ultimate objective of predicting the system. To explain the logic, consider the ideal case that a model essentially captures the behavior of the system. In this case, $\square(S|s_1{=}l)$ would approach $F(S|s_1{=}l)$ as n_w becomes very large. This would lead to $JSD[F(S|s_1{=}l),\square\square(S|s_1{=}l)]$ tending to 0. The key difficulty with this statement is that n_w cannot really be made to approach infinity, which is likely to be necessary to fully confirm a possible equality between $F(S|s_1{=}l)$ and$\square\square(S|s_1{=}l)$. Instead, a realistic optimal level of agreement between $F(S|s_1{=}l)$ and$\square\square(S|s_1{=}l)$, given that we can only do a finite number of walks n_w to create a $\square(S|s_1{=}l)$, would be signalled by the fact that $F(S|s_1{=}l)$ would be indistinguishable from any one of the $n_\square$ realizations $\square_r(S|s_1{=}l)$. In this case, the probability distributions of $m_r(l)$ and $t_{r,r'}(l)$ should be indistinguishable (i.e., should overlap). This implies that the samples of $t_{r,r'}(l)$ act as a baseline check, to see how far the model is from the "ideal" modeling of the system. Figure 1 and Figure 6 show these results for two sample units and occupational series. For most l in the system, the results show that the models are not a perfect match with the system, but the very small values of *JSD* indicate that they are also not that far off.

The use of the $m_r(l)$ samples, as briefly indicated above, allows performance comparison between the models. This is done simply by determining which model leads to a distribution of $m_r(l)$ with *smaller* values. As we see in Figure 1 and Figure 6, the model with memory indeed performs better.

References

Abbott, A., & Hrycak, A. (1990). Measuring resemblance in sequence data: An optimal matching analysis of musicians' careers. *American Journal of Sociology*, *96*(1), 144–85.

Aisenbrey, S., & Fasang, A. E. (2010). New life for old ideas: The "second wave" of sequence analysis bringing the "course" back into the life course. *Sociological Methods & Research*, *38*(3), 420–462.

Axtell, R. L., Guerrero, O. A., & López, E. (2019). Frictional unemployment on labor flow networks. *Journal of Economic Behavior & Organization*, *160*, 184–201.

Barabási, A-L. (2014). *Linked: How everything is connected to everything else and what it means for business, science, and everyday life*. Basic Books.

Bidwell, M. (2017). Managing talent flows through internal and external labor markets. In *The Oxford Handbook of Talent Management* (pp. 281–298). Oxford University Press.

Collet, F., & Hedström, P. (2013). Old friends and new acquaintances: Tie formation mechanisms in an interorganizational network generated by employee mobility. *Social Networks*, *35*(3), 288–299.

Cornwell, B. (2015). *Social sequence analysis: Methods and applications*. Cambridge University Press.

De Feyter, T., & Guerry, M.-A. (2011). Markov models in manpower planning: A review. In *Handbook of Optimization Theory* (pp. 67–88). Nova Science Publishers Inc.

Ginsberg, R. B. (1971). Semi-Markov processes and mobility. *Journal of Mathematical Sociology*, *1*, 233–262.

Guerrero, O. A., & Axtell, R. L. (2013). Employment growth through labor flow networks. *PlosONE*, *8*(5), e60808. https://doi.org/10.1371/journal.pone.0060808

Guerrero, O. A., & López, E. (2015). Firm-to-firm labor flows and the aggregate matching function: A network-based test using employer-employee matched records. *Economic Letters*, *136*, 9–12.

Gunz, H., Lazarova, M., & Mayrhofer, W. (2020). *The Routledge companion to career studies*. Routledge.

Lin, J. (1991). Divergence measures based on Shannon entropy. *IEEE Transactions on Information Theory*, *37*(1), 145–151.

López, E., Guerrero, O. A., & Axtell, R. L. (2020). A network theory of inter-firm labor flows. *EPJ Data Science*, *9*, 33.

Mattis, J. (2018). *Summary of the 2018 National Defense Strategy of the United States of America*. Joint Chiefs of Staff.

Mealy, P., del Rio-Chanona, R. M., & Farmer, J. D. (2018). *What you do at work matters: New lens on labour*. SSRN. http://dx.doi.org/10.2139/ssrn.3143064

Newman, M. (2018). *Networks* (2nd ed.). Oxford University Press.

Rosvall, M., Esquivel, A. V., Lancichinetti, A., West, J. D., & Lambiotte, R. (2014). Memory in network flows and its effects on spreading dynamics and community detection. *Nature Communications*, *5*, 4630.

Secretary of the Army. (2019). *The Army people strategy*. Office of the Assistant Secretary of the Army, Manpower, and Reserve Affairs.

Stewman, S. (1986). Demographic models of internal labor markets. *Administrative Science Quarterly*, *31*(2), 212–247.

Stewman, S., & Konda, S. L. (1983). Careers and organizational labor markets: Demographic models of organizational behavior. *American Journal of Sociology*, *88*(4), 637– 685.

Stewman, S., & Yeh, K. S. (1991). Structural pathways and switching mechanisms for individual careers. *Research in Social Stratification and Mobility*, *10*, 133–168.

Wang, J. (2005, February). *A review of operations research applications in workforce planning and potential modelling of military training*. DSTO Systems Sciences Laboratory.

White, H. C. (1970). *Chains of opportunity: System models of mobility in organizations*. Harvard University Press.

Acquisition Research Program
Naval Postgraduate School
555 Dyer Road, Ingersoll Hall
Monterey, CA 93943

www.acquisitionresearch.net