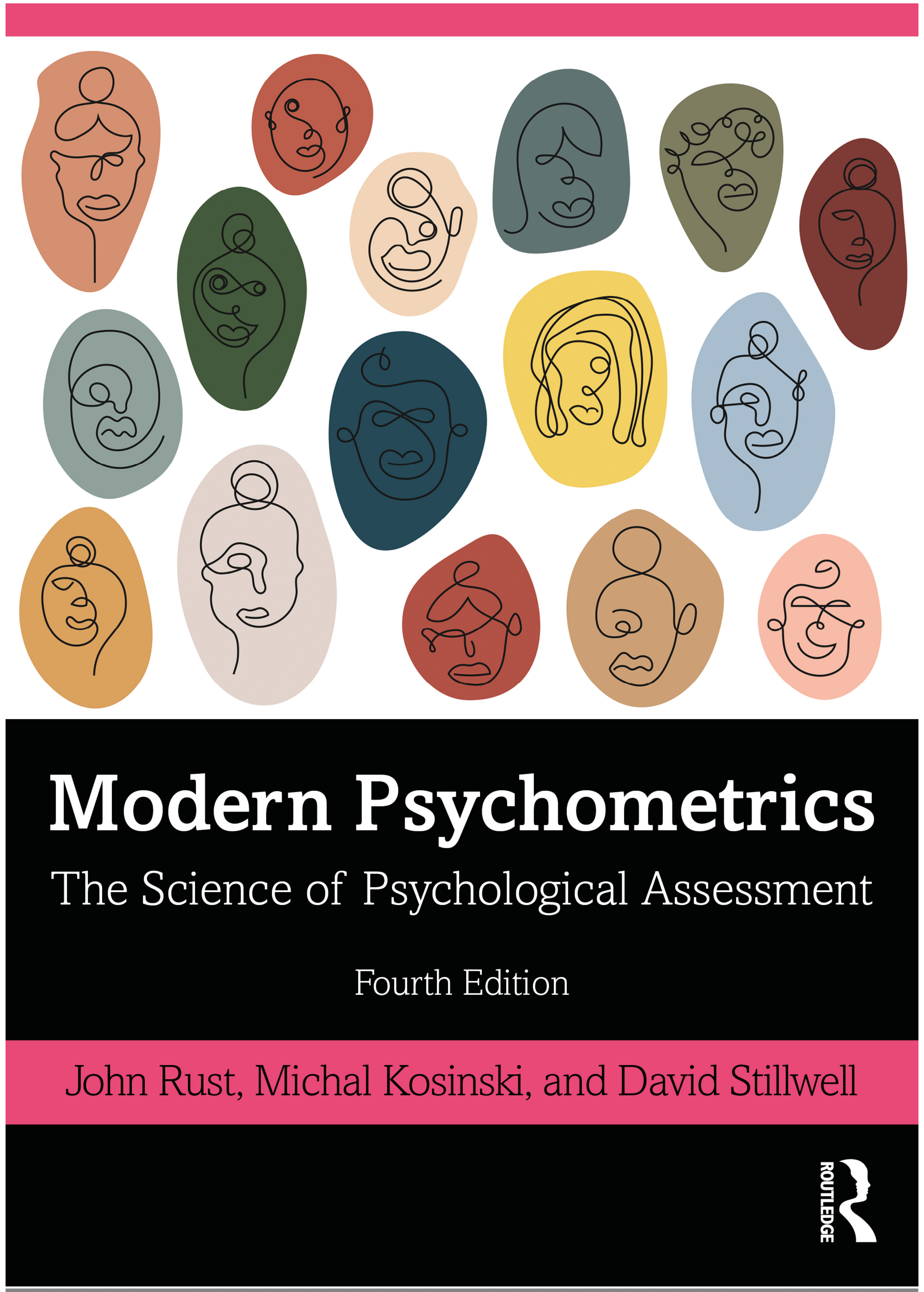
Discussão dos textos

Employing digital footprints in psychometrics
Education Data Science- Past, Present, Future

Prof. Dr. Ricardo Primi

UNIVERSIDADE SÃO FRANCISCO



8 Employing digital footprints in psychometrics

Introduction

Decades of research and practical applications have shown that well-made tests and self-reported questionnaires can be reliable, practical, and accurate. They have been successfully applied across diverse contexts, ranging from recruitment and high-stakes educational assessments to clinical diagnosis. New methodologies such as computer adaptive testing (see Chapter 5) and item generators (see Chapter 9) are becoming more widespread, further driving the quality of such assessments.

At the same time, however, there are major flaws inherent to self-reported questions and test items. First is their temporal character and low ecological validity: assessments offer only a brief window into respondents' opinions and performance, and they are often administered in an artificial environment, such as in an assessment center and under time pressure. During the brief interaction with a questionnaire or test, a respondent may be affected by factors such as the testing environment, stress, fatigue, or even the weather. Consequently, their scores reflect not only the traits being measured but also these external factors, decreasing the measurement's validity and reliability.

Second, traditional assessments are limited to capturing respondents' explicit, conscious, and motivated opinions and behaviors. Consequently, they are vulnerable to cheating and misrepresentation, particularly when much depends on the scores, such as in the context of recruitment or entrance exams. Misrepresentation is often unconscious, driven by a wide range of unconscious cognitive biases. Availability bias, for instance, leads to overestimating the frequency of thoughts or behaviors that are easily accessible in one's memory. For example, after weeks spent preparing for a job interview, job candidates are likely to underestimate how social they normally are. Another common bias, the reference-group bias, describes the difficulty of comparing one's trait levels with the average in a general population. Instead, we tend to compare ourselves with those around us, or a reference group. An extraverted actor, for instance, might genuinely believe themselves to be introverted if they are surrounded by even more extraverted peers. Consequently, even widely used and well-validated assessments are often relatively poor predictors of many basic real-life outcomes such as performance at work, well-being, or physical activity.

How can we solve these and other limitations of traditional assessments? One approach would be to replace tests and questionnaires—narrow snapshots of respondents' self-reported behaviors—with long-term observations of actual behaviors, preferences, and performance in the natural environment. One could follow the respondents around for, let us say, a full year, meticulously recording all the times when they expressed

Síntese

Limitações de instrumentos de auto relato

Alternativa: "*long-term observations of actual behaviors, preferences, and performance in the natural environment*"

- *Our ongoing migration to the digital environment opened up a myriad of ways in which our behaviors, preferences, and performance can be recorded in an unobtrusive, cheap, and convenient way.*
- *web-browsing logs, records of transactions from online and off-line marketplaces, photos and videos, GPS location logs, media playlists, voice and video call logs, language used in tweets or emails, and much more.*

Digital footprints

Tipos de dados

Síntese

Aplicações

substitutos de medidas tradicionais, novos contextos e novas medidas (sistemas de recomendação), predição, estudo do comportamento humano, dar suporte a medidas existentes

Vantagens e desafios

validade ecológica, detalhamento e longitude, menor controle da situação de testagem, velocidade e nao-intrusão, privacidade, ausência de anonimato, viés, enriquecimento de avaliaçào de construtor psicológicos

Desenvolvendo medidas

- Dados
  - feedback, representatividade, N
- Análises
  - Matriz de dados
  - Matriz gigante e esparsa
  - Redução de dimensionalidade (SVD, LDA)

Original Matrix

	google.com	facebook.com	etsy.com	deviantart.com	rottentomatoes.com	imdb.com	vimeo.com
Jason	13	9					
Neville	30	25	6		7		
David	14	13	17	21			
Susan	5		27	33			
Stan			17	15			
Michal		4	14	16	22	34	
John	6				12	11	
Khalifa			3				5
Sara							11

Trimming →

Trimmed Matrix

	google.com	facebook.com	etsy.com	deviantart.com	rottentomatoes.com	imdb.com
Jason	13	9				
Neville	30	25	6		7	
David	14	13	17	21		
Susan	5		27	33		
Stan			17	15		
Michal		4	14	16	22	34
John	6				12	11

Figure 8.1 A hypothetical respondent-footprint matrix representing the frequencies of website visits and its trimmed version (see text for details). Cells represent the number of times a given respondent visited a given website. Shading based on the frequency was added to enhance readability. Zeros were removed for clarity.

Popular languages for statistical programming (e.g., Python, R, and MATLAB) provide off-the-shelf functions that allow for reducing matrix dimensionality using SVD. To preserve computational resources, make sure to use a sparse SVD function, or a function that can take a sparse matrix as an input without converting it into a nonsparse format. SVD decomposes a matrix into three matrices (**U**, **V**, and **Σ**) exposing its underlying structure. Matrices **U** and **V** contain singular vectors subsuming the patterns present in the original matrix. Diagonal matrix **Σ** contains singular values representing the importance of each of the singular vectors. (A diagonal matrix is a matrix where only the diagonal cells are filled with values.) The product of **U**, **Σ**, and transposed **V** ($\mathbf{U\ \Sigma\ V^T}$) is equal to the original matrix.

The first singular vector subsumes the most prominent pattern in the matrix, and the subsequent vectors represent patterns of decreasing importance. Thus, the dimensionality of the matrix can be reduced by discarding some of the less important singular vectors. The product of the resulting trimmed matrices **U**, **Σ**, and $\mathbf{V^T}$ does not represent the original matrix exactly, but provides its approximation.

Respondents (U)

	SVD1	SVD2	SVD3
Jason	.10	.33	.05
Neville	.32	.79	.19
David	.46	.20	-.24
Susan	.57	-.23	-.48
Stan	.29	-.16	-.26
Michal	.51	-.39	.68
John	.12	.00	.37

Footprints (V)

	SVD1	SVD2	SVD3
google.com	.32	.70	.07
facebook.com	.26	.56	.12
etsy.com	.56	-.15	-.27
deviantart.com	.62	-.28	-.35
rottentomatoes.com	.23	-.07	.54
imdb.com	.28	-.31	.70

Figure 8.3 Respondents' (matrix V) and websites' (matrix U) scores on three singular vectors extracted from the trimmed respondent-footprint matrix presented in Figure 8.1.

Respondents ($U_{varimax}$)

	$SVD_{varimax}1$	$SVD_{varimax}2$	$SVD_{varimax}3$
Jason	.00	.37	.00
Neville	.06	.93	.12
David	.41	.44	.00
Susan	.65	.08	-.01
Stan	.34	.00	.00
Michal	.33	.04	1.04
John	.00	.11	.41

Footprints ($V_{varimax}$)

	$SVD_{varimax}1$	$SVD_{varimax}2$	$SVD_{varimax}3$
google.com	.02	.77	-.02
facebook.com	-.01	.62	.04
etsy.com	.64	.06	.00
deviantart.com	.77	-.06	-.01
rottentomatoes.com	-.02	.09	.58
imdb.com	.02	-.08	.81

Figure 8.4 Varimax-rotated singular vectors from Figure 8.3.

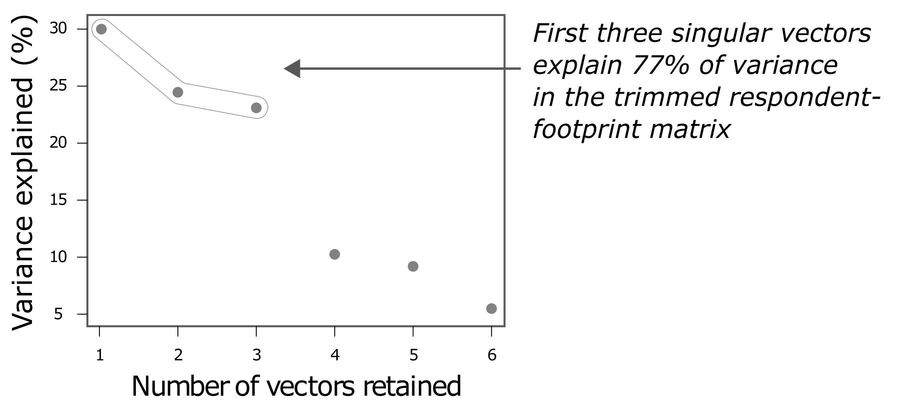


Figure 8.2 Variance explained by consecutive singular vectors in the trimmed respondent-footprint matrix presented in Figure 8.1 (right panel).

Figure 8.6 presents two matrices produced by LDA expressing the associations between clusters and websites (matrix $\boldsymbol{\beta}$) and clusters and respondents (matrix $\boldsymbol{\gamma}$). Matrix $\boldsymbol{\beta}$ contains probabilities of a particular website being visited (relatively, within a cluster). For example, take cluster LDA1, which groups art-related websites. When a participant visits a website in this cluster, they will pick Deviantart.com with a probability of .53, Etsy.com with a probability of .46, and Google.com with a probability of .01. Note that the probabilities sum to 1 in each column. This is because we are dealing with mutually exclusive events encompassing all possible outcomes: if a participant visits a given cluster, they must choose one of the websites in the matrix.

Matrix $\boldsymbol{\gamma}$ contains probabilities of a respondent visiting one of the websites in a given cluster. For example, David's probability of visiting websites in cluster LDA1 (Etsy.com and Deviantart.com) equals .6, while his probability of visiting websites in cluster LDA2 (Google.com and Facebook.com) equals .4. Compare this with the respondent-footprint matrix (Figure 8.1) showing that David visited exclusively websites in those clusters, and that he was more likely to visit those belonging to LDA1. In matrix $\boldsymbol{\gamma}$, the probabilities

	LDA1	LDA2	LDA3
Jason	.00	.99	.00
Neville	.00	1.0	.00
David	.60	.40	.00
Susan	.92	.08	.00
Stan	.99	.00	.00
Michal	.33	.00	.67
John	.00	.28	.72

Respondents (γ)

	LDA1	LDA2	LDA3
google.com	.01	.52	.00
facebook.com	.00	.36	.05
etsy.com	.46	.05	.00
deviantart.com	.53	.00	.00
rottentomatoes.com	.00	.07	.39
imdb.com	.00	.00	.55

Footprints (β)

Figure 8.6 Three LDA topics extracted from the trimmed respondent-footprint matrix presented in Figure 8.1. Matrix γ shows associations between respondents and clusters; matrix β shows the associations between websites and clusters.

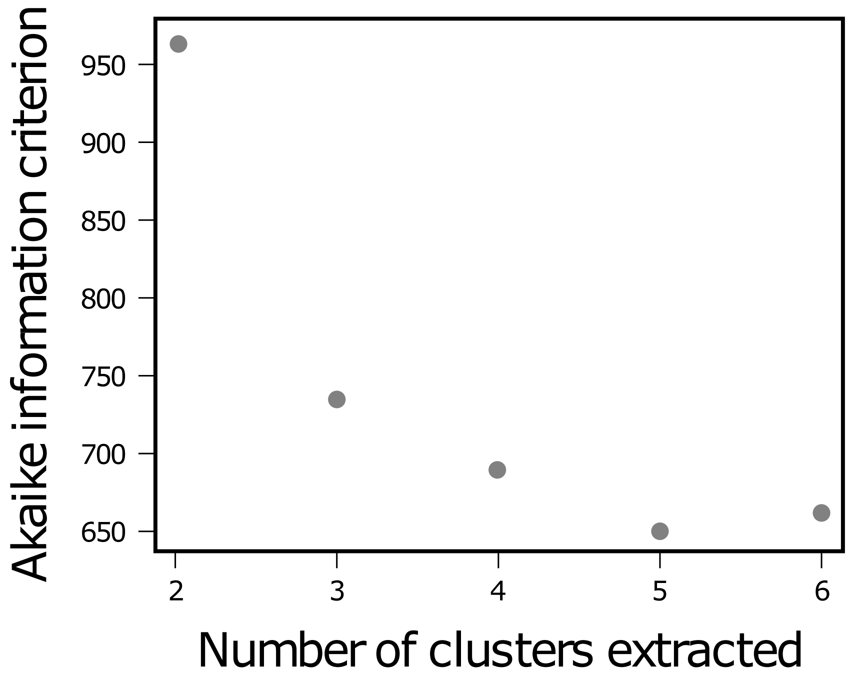


Figure 8.5 AIC and the number of LDA clusters extracted from the trimmed respondent-footprint matrix presented in Figure 8.1. (Note that the minimum number of clusters that can be extracted is k = 2.)

Desenvolvendo modelos preditivos

1. Outcome
2. Dividir amostra em treino / teste
3. Redução de dimensionalidade
4. Escores sujeitos (matriz gama ou U) e treino do modelo
5. Previsão na base teste
6. Validade e precisão

AERA Open
January-December 2021, Vol. 7, No. 1, pp. 1–12
DOI: 10.1177/23328584211052055
Article reuse guidelines: sagepub.com/journals-permissions

Education Data Science: Past, Present, Future

Daniel A. McFarland
Saurabh Khanna
Benjamin W. Domingue
Stanford University
Zachary A. Pardos
University of California, Berkeley

This AERA Open special topic concerns the large emerging research area of education data science (EDS). In a narrow sense, EDS applies statistics and computational techniques to educational phenomena and questions. In a broader sense, it is an umbrella for a fleet of new computational techniques being used to identify new forms of data, measures, descriptives, predictions, and experiments in education. Not only are old research questions being analyzed in new ways but also new questions are emerging based on novel data and discoveries from EDS techniques. This overview defines the emerging field of education data science and discusses 12 articles that illustrate an AERA-angle on EDS. Our overview relates a variety of promises EDS poses for the field of education as well as the areas where EDS scholars could successfully focus going forward.

Keywords: *data science, network analysis, natural language processing, machine learning, learning analytics, data mining*

Síntese

“data science” as an alias for computer science and statistics respectively

International Association for Statistical Computing (IASC) in 1977 was established with a “mission to link traditional statistical methodology, modern computer technology, and the knowledge of domain experts in order to convert data into information and knowledge.”

Data science had risen to occupy a unique disciplinary position on account of (a) being more application oriented as it targets solutions to real-world challenges (Donoho, 2017), (b) coupling quantitative and qualitative research across disciplines (Dhar, 2013), and (c) being largely focused on digital structured and unstructured data (Silver, 2020).

Learning Analytics, Machine Learning, Artificial Intelligence, Data Science, and Natural Language Processing.

Tópicos

In a narrow sense, one could conceptualize EDS as the application of tools and perspectives from statistics and computer science to educational phenomena and problems.

But we argue for a more expansive definition where EDS is an umbrella for a range of new and often nontraditional quantitative methods (such as machine learning, network analysis, and natural language processing) applied to educa- tional problems often using novel data.

Novos dados: dados de nível micro + aspecto temporal, nível meso, nível macro

Novos métodos: textos

NLP

Natural language processing (NLP), as a subdomain of machine learning, warrants particular attention. It has been extensively applied to text data in education settings—either collected firsthand or transcribed from media recordings.

Broadly, NLP techniques can be applied to large text corpora in education to understand traits like sentiment embedded in the text, the novelty in information presented, and topics identified by topic modeling and topic classification approaches (Islam et al., 2012; Lucy et al., 2020).

In a classroom setting, NLP algorithms can dynamically assess reading proficiency for a student and generate real-time feedback for improvement (Li et al., 2017). Modern NLP algorithms are being used to provide actionable feedback around prose, grammar, and general writing mechanics (Alhawiti, 2014; Shum et al., 2016) and to allow for examinations of, for example, the potential signature of class in written elements of educational materials (Alvero et al., 2021). In addition to a student-facing component, NLP platforms can provide a teacher-facing component as well. This helps enable teachers conduct robust formative assessments that might otherwise be difficult in classrooms with large student-teacher ratios (Burstein et al., 2014; Chapelle & Chung, 2010).

Social network analysis (node2vec)

Integration in emerging social networks explains academic failure and success

Christoph Stadtfeld[a,1,2], András Vörös[a,1], Timon Elmer[a], Zsófia Boda[a], and Isabel J. Raabe[a,b]

[a]Chair of Social Networks, Department of Humanities, Social and Political Sciences, ETH Zürich, 8092 Zürich, Switzerland; and [b]Institute of Sociology, University of Zürich, 8006 Zürich, Switzerland

Edited by H. Russell Bernard, University of Florida, Gainesville, FL, and approved November 16, 2018 (received for review July 2, 2018)

Academic success of students has been explained with a variety of individual and socioeconomic factors. Social networks that informally emerge within student communities can have an additional effect on their achievement. However, this effect of social ties is difficult to measure and quantify, because social networks are multidimensional and dynamically evolving within the educational context. We repeatedly surveyed a cohort of 226 engineering undergraduates between their first day at university and a crucial examination at the end of the academic year. We investigate how social networks emerge between previously unacquainted students and how integration in these networks explains academic success. Our study measures multiple important dimensions of social ties between students: their positive interactions, friendships, and studying relations. By using statistical models for dynamic network data, we are able to investigate the processes of social network formation in the cohort. We find that friendship ties informally evolve into studying relationships over the academic year. This process is crucial, as studying together with others, in turn, has a strong impact on students' success at the examination. The results are robust to individual differences in socioeconomic background factors and to various indirect measures of cognitive abilities, such as prior academic achievement and being perceived as smart by other students. The findings underline the importance of understanding social network dynamics in educational settings. They call for the creation of university environments promoting the development of positive relationships in pursuit of academic success.

social networks | sociology of education | network dynamics |

Social networks are dynamic and complex systems (10–13); therefore, their impact on academic achievement is difficult to study. We argue that it is not sufficient to assess students' integration in social networks in critical periods (e.g., when they prepare for a crucial examination), but it is equally important to understand how those networks evolved in the months and years before. The social networks between students in this study were basically nonexistent before their first university day, as 81% of them had just moved from other cities or countries. By the end of the first year, however, 549 friendship ties emerged, and 61% of the participating students reported that one of their best friends in life was among their fellow students.

To capture the dynamic nature of social networks, we conducted multiple surveys throughout the academic year, measuring three dimensions of student relations: positive interaction, friendship, and studying together. We further collected self-reported measures of socioeconomic background (sex, age, first language, parental education, and support ties outside the university), abilities and motivation [high-school grade point average (GPA), study motivation, and time spent on studying], psychological well-being (stress, depression, and anxiety), and different types of peer perceptions (being perceived as smart) and behaviors (having a side job and free-time activities). Our study builds upon prior educational research in different scientific fields that looked into the relationship between social integration and academic performance. Prior work, for example, investigated the role of aggregated peer effects (14, 15),

Significance

SOCIAL SCIENCES

node2vec: Scalable Feature Learning for Networks

Aditya Grover
Stanford University
adityag@cs.stanford.edu

Jure Leskovec
Stanford University
jure@cs.stanford.edu

ABSTRACT

Prediction tasks over nodes and edges in networks require careful effort in engineering features used by learning algorithms. Recent research in the broader field of representation learning has led to significant progress in automating prediction by learning the features themselves. However, present feature learning approaches are not expressive enough to capture the diversity of connectivity patterns observed in networks.

Here we propose *node2vec*, an algorithmic framework for learning continuous feature representations for nodes in networks. In *node2vec*, we learn a mapping of nodes to a low-dimensional space of features that maximizes the likelihood of preserving network neighborhoods of nodes. We define a flexible notion of a node's network neighborhood and design a biased random walk procedure, which efficiently explores diverse neighborhoods. Our algorithm generalizes prior work which is based on rigid notions of network neighborhoods, and we argue that the added flexibility in exploring neighborhoods is the key to learning richer representations.

We demonstrate the efficacy of *node2vec* over existing state-of-the-art techniques on multi-label classification and link prediction in several real-world networks from diverse domains. Taken together, our work represents a new way for efficiently learning state-of-the-art task-independent representations in complex networks.

Categories and Subject Descriptors: H.2.8 **[Database Management]**: Database applications—*Data mining*; I.2.6 **[Artificial Intelligence]**: Learning

General Terms: Algorithms; Experimentation.

Keywords: Information networks, Feature learning, Node embeddings, Graph representations.

1. INTRODUCTION

Many important tasks in network analysis involve predictions over nodes and edges. In a typical node classification task, we are interested in predicting the most probable labels of nodes in a network [33]. For example, in a social network, we might be interested in predicting interests of users, or in a protein-protein interaction network we might be interested in predicting functional labels of proteins [25, 37]. Similarly, in link prediction, we wish to predict whether a pair of nodes in a network should have an edge connecting them [18]. Link prediction is useful in a wide variety of domains; for instance, in genomics, it helps us discover novel interactions between genes, and in social networks, it can identify real-world friends [2, 34].

Any supervised machine learning algorithm requires a set of informative, discriminating, and independent features. In prediction problems on networks this means that one has to construct a feature vector representation for the nodes and edges. A typical solution involves hand-engineering domain-specific features based on expert knowledge. Even if one discounts the tedious effort required for feature engineering, such features are usually designed for specific tasks and do not generalize across different prediction tasks.

An alternative approach is to *learn* feature representations by solving an optimization problem [4]. The challenge in feature learning is defining an objective function, which involves a trade-off in balancing computational efficiency and predictive accuracy. On one side of the spectrum, one could directly aim to find a feature representation that optimizes performance of a downstream prediction task. While this supervised procedure results in good accuracy, it comes at the cost of high training time complexity due to a blowup in the number of parameters that need to be estimated. At the other extreme, the objective function can be defined to be independent of the downstream prediction task and the representations can be learned in a purely unsupervised way. This makes the optimization computationally efficient and with a carefully designed objective, it results in task-independent features that closely match task-specific approaches in predictive accuracy [21, 23].

However, current techniques fail to satisfactorily define and optimize a reasonable objective required for scalable unsupervised feature learning in networks. Classic approaches based on linear and non-linear dimensionality reduction techniques such as Principal Component Analysis, Multi-Dimensional Scaling and their extensions [3, 27, 30, 35] optimize an objective that transforms a representative data matrix of the network such that it maximizes the variance of the data representation. Consequently, these approaches invariably involve eigendecomposition of the appropriate data matrix which is expensive for large real-world networks. Moreover, the resulting latent representations give poor performance on various prediction tasks over networks.

Alternatively, we can design an objective that seeks to preserve local neighborhoods of nodes. The objective can be efficiently optimized using stochastic gradient descent (SGD) akin to backpropogation on just single hidden-layer feedforward neural networks. Recent attempts in this direction [24, 28] propose efficient algorithms but rely on a rigid notion of a network neighborhood, which results in these approaches being largely insensitive to connectivity patterns unique to networks. Specifically, nodes in networks

KDD '16, August 13 - 17, 2016, San Francisco, CA, USA

ISBN 978-1-4503-4232-2/16/08…$15.00

DOI: http://dx.doi.org/10.1145/2939672.2939754

Package ‘node2vec’

January 14, 2021

Title Algorithmic Framework for Representational Learning on Graphs

Version 0.1.0

Description Given any graph, the 'node2vec' algorithm can learn continuous feature representations for the nodes, which can then be used for various downstream machine learning tasks.The techniques are detailed in the paper ``node2vec: Scalable Feature Learning for Networks'' by Aditya Grover, Jure Leskovec(2016),available at <arXiv:1607.00653>.

License GPL (>= 3)

Encoding UTF-8

LazyData true

RoxygenNote 7.1.0

Imports data.table, igraph, word2vec, rlist, dplyr, vctrs, vegan

Depends R (>= 2.10)

NeedsCompilation no

Author Yang Tian [aut, cre],
Xu Li [aut],
Jing Ren [aut]

Maintainer Yang Tian <tianyang1211@126.com>

Repository CRAN

Date/Publication 2021-01-14 09:00:02 UTC

R topics documented:

Discussão

Modeling

In that project, the addition of rich data and sophisticated modeling techniques did not substantially increase the predictability of several life course outcomes of relevance in the study of young people. We think these results are useful in terms of setting expectations: Behavioral science in general and educational science in particular are challenging. Most innovations on the data or computational side should be anticipated to bring only marginal improve- ments in our understanding.

Bias

4 pontos (p9): Amostra enviesada, natureza da predição (projetar passado no futuro), consequência da velocidade, falta de supervisão humana

Check for updates

Measuring the predictability of life outcomes with a scientific mass collaboration

Matthew J. Salganik[a,1], Ian Lundberg[a], Alexander T. Kindel[a], Caitlin E. Ahearn[b], Khaled Al-Ghoneim[c], Abdullah Almaatouq[d,e], Drew M. Altschul[f], Jennie E. Brand[b,g], Nicole Bohme Carnegie[h], Ryan James Compton[i], Debanjan Datta[j], Thomas Davidson[k], Anna Filippova[l], Connor Gilroy[m], Brian J. Goode[n], Eaman Jahani[o], Ridhi Kashyap[p,q,r], Antje Kirchner[s], Stephen McKay[t], Allison C. Morgan[u], Alex Pentland[e], Kivan Polimis[v], Louis Raes[w], Daniel E. Rigobon[x], Claudia V. Roberts[y], Diana M. Stanescu[z], Yoshihiko Suhara[e], Adaner Usmani[aa], Erik H. Wang[z], Muna Adem[bb], Abdulla Alhajri[cc], Bedoor AlShebli[dd], Redwane Amin[ee], Ryan B. Amos[y], Lisa P. Argyle[ff], Livia Baer-Bositis[gg], Moritz Büchi[hh], Bo-Ryehn Chung[ii], William Eggert[jj], Gregory Faletto[kk], Zhilin Fan[ll], Jeremy Freese[gg], Tejomay Gadgil[mm], Josh Gagné[gg], Yue Gao[nn], Andrew Halpern-Manners[bb], Sonia P. Hashim[y], Sonia Hausen[gg], Guanhua He[oo], Kimberly Higuera[gg], Bernie Hogan[pp], Ilana M. Horwitz[qq], Lisa M. Hummel[gg], Naman Jain[x], Kun Jin[rr], David Jurgens[ss], Patrick Kaminski[bb,tt], Areg Karapetyan[uu,vv], E. H. Kim[gg], Ben Leizman[y], Naijia Liu[z], Malte Möser[y], Andrew E. Mack[z], Mayank Mahajan[y], Noah Mandell[ww], Helge Marahrens[bb], Diana Mercado-Garcia[qq], Viola Mocz[xx], Katariina Mueller-Gastell[gg], Ahmed Musse[yy], Qiankun Niu[ee], William Nowak[zz], Hamidreza Omidvar[aaa], Andrew Or[y], Karen Ouyang[y], Katy M. Pinto[bbb], Ethan Porter[ccc], Kristin E. Porter[ddd], Crystal Qian[y], Tamkinat Rauf[gg], Anahit Sargsyan[eee], Thomas Schaffner[y], Landon Schnabel[gg], Bryan Schonfeld[z], Ben Sender[fff], Jonathan D. Tang[y], Emma Tsurkov[gg], Austin van Loon[gg], Onur Varol[ggg,hhh], Xiafei Wang[iii], Zhi Wang[hhh,jjj], Julia Wang[y], Flora Wang[fff], Samantha Weissman[y], Kirstie Whitaker[kkk,lll], Maria K. Wolters[mmm], Wei Lee Woon[nnn], James Wu[ooo], Catherine Wu[y], Kengran Yang[aaa], Jingwen Yin[ll], Bingyu Zhao[ppp], Chenyun Zhu[ll], Jeanne Brooks-Gunn[qqq,rrr], Barbara E. Engelhardt[y,ii], Moritz Hardt[sss], Dean Knox[z], Karen Levy[ttt], Arvind Narayanan[y], Brandon M. Stewart[a], Duncan J. Watts[uuu,vvv,www], and Sara McLanahan[a,1]

Contributed by Sara McLanahan, January 24, 2020 (sent for review October 1, 2019; reviewed by Sendhil Mullainathan and Brian Uzzi)

How predictable are life trajectories? We investigated this question with a scientific mass collaboration using the common task method; 160 teams built predictive models for six life outcomes using data from the Fragile Families and Child Wellbeing Study, a high-quality birth cohort study. Despite using a rich dataset and applying machine-learning methods optimized for prediction, the best predictions were not very accurate and were only slightly better than those from a simple benchmark model. Within each outcome, prediction error was strongly associated with the family being predicted and weakly associated with the technique used to generate the prediction. Overall, these results suggest practical limits to the predictability of life outcomes in some settings and illustrate the value of mass collaborations in the social sciences.

life course | prediction | machine learning | mass collaboration

Social scientists studying the life course have described social patterns, theorized factors that shape outcomes, and estimated causal effects. Although this research has advanced scientific understanding and informed policy interventions, it is unclear how much it translates into an ability to predict individual life outcomes. Assessing predictability is important for three reasons. First, accurate predictions can be used to target assistance to children and families at risk (1, 2). Second, predictability of a life outcome from a person's life trajectory can indicate social rigidity (3), and efforts to understand differences in predictability across social contexts can stimulate scientific discovery and improve policy-making (4). Finally, efforts to improve predictive performance can spark developments in theory and methods (5).

In order to measure the predictability of life outcomes for children, parents, and households, we created a scientific mass collaboration. Our mass collaboration—the Fragile Families Challenge—used a research design common in machine learning but not yet common in the social sciences: the common task method (6). To create a project using the common task method, an organizer designs a prediction task and then recruits a large, diverse group of researchers who complete the task by predicting the exact same outcomes using the exact same data. These predictions are then evaluated with the exact same error metric that exclusively assesses their ability to predict held-out data: data that are held by the organizer and not available to participants. Although the structure of the prediction task is completely standardized, participants are free to use any technique to generate predictions.

The common task method produces credible estimates of predictability because of its design. If predictability is higher than expected, the results cannot be dismissed because of concerns about overfitting (7) or researcher degrees of freedom (8). Alternatively, if predictability is lower than expected, the results cannot be dismissed because of concerns about the limitations

Author contributions: M.J.S., I.L., A.T.K., J.B.-G., B.E.E., M.H., K.L., A.N., B.M.S., D.J.W., and S. McLanahan designed research; M.J.S., I.L., A.T.K., C.E.A., K.A.-G., A. Almaatouq, D.M.A., J.E.B., N.B.C., R.J.C., D.D., T.D., A.F., C.G., B.J.G., E.J., R.K., A. Kirchner, S. McKay, A.C.M., A.P., K.P., L.R., D.E.R., C.V.R., D.M.S., Y.S., A.U., E.H.W., M.A., A. Alhajri, B.A., R.A., R.B.A., L.P.A., L.B.-B., M.B., B.-R.C., W.E., G.F., Z.F., J.F., T.G., J.G., Y.G., A.H.-M., S.P.H., S.H., G.H., K.H., B.H., I.M.H., L.M.H., N.J., K.J., D.J., P.K., A. Karapetyan, E.H.K., B.L., N.L., M. Möser, A.E.M., M. Mahajan, N.M., H.M., D.M.-G., V.M., K.M.-G., A.M., Q.N., W.N., H.O., A.O., K.O., K.M.P., E.P., K.E.P., C.Q., T.R., A.S., T.S., L.S., B. Schonfeld, B. Sender, J.D.T., E.T., A.v.L., O.V., X.W., Z.W., J. Wang, F.W., S.W., K.W., M.K.W., W.L.W., J. Wu, C.W., K.Y., J.Y., B.Z., and C.Z. analyzed data; and M.J.S., I.L., A.T.K., D.K., and S. McLanahan wrote the paper.

Reviewers: S.M., University of Chicago; and B.U., Northwestern University.

Competing interest statement: B.E.E. is on the scientific advisory boards of Celsius Therapeutics and Freenome, is currently employed by Genomics plc and Freenome, and is on a year leave-of-absence from Princeton University.

Data deposition: Predictions, code, and narrative explanations for valid submissions to the Fragile Families Challenge and code to reproduce the results of this paper are available from Dataverse at https://doi.org/10.7910/DVN/CXSECU. Data used in the Fragile Families Challenge are currently available to approved researchers from the Princeton University Office of Population Research Data Archive at https://opr.princeton.edu/archive/.

See online for related content such as Commentaries.

[1]To whom correspondence may be addressed. Email: mjs3@princeton.edu or mclanaha@princeton.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.1915006117/-/DCSupplemental.

First published March 30, 2020.

https://fragilefamilies.princeton.edu

https://www.fragilefamilieschallenge.org