From:	Garrett Doherty
To:	Mabry, Patricia L
Cc:	brian.c.martinson@gmail.com; Liu, Xiaozhong; Valente, Thomas; Haven, Alyse L
Subject:	Re: Seeking info re: accepted panel for ASIST AM22
Date:	Wednesday, August 10, 2022 3:33:22 PM
Attachments:	Mabry P Panel22.pdf

Dear Patricia Mabry,

Thank you for writing in about the accepted panel.

I can confirm that this Panel has been accepted for presentation at the 85th Annual Meeting of the Association for Information Science & Technology | Oct. 29 – Nov. 1, 2022 | Pittsburgh, PA.

"Exploiting Interdisciplinary Research Design for Temporally Complex Big Data: Discussion of a Case-Study Using on Heterogenous Bibliographic Big Data"
by
Mabry, Patricia L., Research Dept., HealthPartners Institute, USA |
Patricia.L.Mabry@HealthPartners.com
Martinson, Brian, Dept of Medicine, University of Minnesota, USA |
Brian.C.Martinson@gmail.com
Valente, Thomas, University of Southern California, Irvine, USA |
Thomas.Valente@med.usc.edu
Xiaozhong Liu, Data Science & Computer Science, Worcester Polytechnic Institute, USA |
xliu14@wpi.edu

The panel paper will be published in the *Proceedings of the Association for Information Science and Technology,* Volume 59, Issue 1.
The proceedings will be published online by October 29, 2022, at:
https://asistdl.onlinelibrary.wiley.com/journal/23739231

Attached is a final PDF proof of the Panel paper.

Best,
Garrett
--
Garrett Doherty
Managing Editor
The Association for Information Science & Technology
JASIS&T | *ARIST* | *Information Matters*
673 Potomac Station Drive, Suite 155, Leesburg, VA 20176
+1 240-670-7827

From: "Mabry, Patricia L" <Patricia.L.Mabry@HealthPartners.Com>
Date: Wednesday, August 10, 2022 at 3:58 PM
To: Garrett Doherty <gdoherty@asist.org>
Cc: "brian.c.martinson@gmail.com" <brian.c.martinson@gmail.com>, "Liu, Xiaozhong" <xliu14@wpi.edu>, "Valente, Thomas" <Thomas.Valente@med.usc.edu>, "Haven, Alyse L" <Alyse.L.Haven@HealthPartners.Com>
Subject: Seeking info re: accepted panel for ASIST AM22

Dear Garrett,

I am preparing my annual progress report to NSF and would like to include this accepted panel presentation and conference paper in the report. I spoke to NSF about this and they requested "proof of acceptance" of the paper for publication in the conference proceedings.

Could you kindly provide the following,:

1. Correspondence to me via Reply All to this message, clearly stating that
    - the panel has been accepted for presentation at the 2022 ASIST Annual Meeting
    - the panel paper will be published in the *Proceedings of the Association for Information Science and Technology* 59(?) along with estimated publication date
2. pdf of the corrected proof (I only have a copy of the uncorrected proof with my mark up, which I previously submitted to you)

Thank you very much!

PM

Patricia L. Mabry, Ph.D. | Research Investigator, Research Department
About me: See ORCID Profile: https://orcid.org/0000-0002-6374-8679
8170 33rd Ave. S. | Mail stop 21112R | PO Box 1524| Minneapolis, MN, 55440-1524
Cell: (703) 303-5628 | Desk: 952-967-6907 | Patricia.L.Mabry@healthpartners.com
HealthPartners Institute | make good happen

From: Garrett Doherty <asist2022@conftool.org>
Sent: Friday, July 1, 2022 2:36 PM
To: Woodrow, Holly G <Holly.G.Woodrow@HealthPartners.Com>
Cc: Mabry, Patricia L <Patricia.L.Mabry@HealthPartners.Com>; Harste, Ann M <Ann.M.Harste@HealthPartners.Com>; brian.c.martinson@gmail.com; xliu14@wpi.edu; Valente, Thomas <Thomas.Valente@med.usc.edu>; Haven, Alyse L <Alyse.L.Haven@HealthPartners.Com>
Subject: [EXTERNAL]Final Review of Panel Paper in ASIST AM22, due July 8, 2022

Dear Patricia Mabry,

Thank you for publishing your Panel Paper submission in the Proceedings of the 2022 ASIST Annual Meeting.

Could you please give a final review of the paper as it will appear in the Proceedings?

At the following link is a PDF of all Short Papers:

https://growthzonesitesprod.azureedge.net/wp-content/uploads/sites/946/PanelsAltEventsJune29.pdf

Please download the file, find your article by clicking on the title in the table of contents, and please review your paper only for any layout errors. Please note that at this final proofing step we cannot revise or edit the text of the paper. Please indicate only those errors made during the layout process of creating this PDF (for example, author names missing from your original submission, or figures or tables accidentally deleted from your original submission).

Please indicate changes by printing those pages which need to be changed, marking the changes on the page, and scanning and returning the pages by e-mail to Garrett Doherty, Managing Editor, at gdoherty@asist.org

Please return changes to files by end of business day, July 8, 2022.

Best,

Garrett

--

Garrett Doherty

Managing Editor

The Association for Information Science & Technology

JASIS&T | ARIST | Information Matters

673 Potomac Station Drive, Suite 155, Leesburg, VA 20176

+1 240-670-7827

--
2022 ASIS&T Annual Meeting
https://www.conftool.org/asist2022/
https://www.asist.org/am22/

Exploiting Interdisciplinary Research Design for Temporally Complex Big Data: Discussion of a Case-Study Using on Heterogenous Bibliographic Big Data

Mabry, Patricia L. Research Dept., HealthPartners Institute, USA | Patricia.L.Mabry@HealthPartners.com

Martinson, Brian Dept of Medicine, University of Minnesota, USA | Brian.C.Martinson@gmail.com

Valente, Thomas University of Southern California, Irvine, USA | Thomas.Valente@med.usc.edu

Xiaozhong Liu Data Science & Computer Science, Worcester Polytechnic Institute, USA | xliu14@wpi.edu

ABSTRACT

Artificial Intelligence (AI) methods are valued for their ability to predict outcomes from dynamically complex data. Despite this virtue, AI is widely criticized as a "black box" i.e., lacking mechanistic explanations to accompany predictions. We introduce a novel interdisciplinary approach that balances the predictive power of data-driven methods with theory-driven explanatory power by presenting a shared use case from four disciplinary perspectives. The use case examines scientific career trajectories through temporally complex, heterogeneous bibliographic big data. Topics addressed include: data representation in complex problems, trade-offs between theoretical, hypothesis-driven, and data-driven approaches, AI trustworthiness, model fairness, algorithm explainability and AI adoption/usability. Panelists and audience members will be prompted to discuss the value of approach presented versus other ways to address the challenges raised by the panel, and to consider their limitations and remaining challenges.

KEYWORDS

Big Data; Interdisciplinary Approach; Heterogenous Bibliographic Big Data

INTRODUCTION

We introduce an innovative, interdisciplinary approach for addressing dynamically complex research questions that combines theory, statistical methods and artificial intelligence. We illustrate with our search for a "Matthew Mechanism" whereby differential accumulation of scholarly social capital leads to greater success for scholars already advantaged in the competition for biomedical research funding. Publication data is used to explore evolution of scholar's research productivity and social capital (via coauthorship) over careers and NIH award data differentiates scholars on funding success. This work presented challenges pertaining to data management, analysis, and interpretation of results arising from the volume of data (including over 34 million publications), heterogeneity of entities to be studied (scholars, publications, journals, NIH awards, research institutions), the multiplicity of relationship types between entities, and temporal dynamics. Panelists from four complementary disciplinary perspectives will engage the audience in a broader discussion of challenges associated with studying complex phenomena in any domain.

SOCIAL NETWORK THEORY

Theory is needed to guide the development of testable hypotheses about the phenomena we study and is integral to the practice of science. In our case study, we seek to understand how differences in social capital accumulation might explain how an outsized proportion of R01 awards increasingly go to a circumscribed group of individuals and institutions. Most scholars studying this funding gap cite the "Matthew Effect" as the cause - shorthand for a process whereby those with resources accumulate ever more resources. In this case we believe some K-awardees are more socially advantaged by dint of their professional connections, which we observe through co-authorship on publications, than others. While NIH awardee data is readily available, NIH applicant data is not. To solve this problem, we study R01 applicants who are in the NIH awards data base by dint of a previous NIH award of another type - the NIH Individual Mentored Career Development Award (MK; K01, K08, K23). The "Mentored K" Award is expressly designed to help awardees gain "research independence" (i.e., Principal Investigator on R01 or equivalent grant). In this presentation, Dr. Valente explains the theoretical underpinnings and rationale for our approach.

The "Matthew Effect" is predicated on social network theory, and specifically on theories of social capital accumulation. Specifically, "Matthew Effect," posits that better-known scientists garner more recognition than the lesser-known peers in spite of similar achievements (Merton 1968, 1988). Evidence is consistent with the notion that small differences between MK awardees get amplified over time, through differential rates of (dis)advantage accrual (Katz & Matter, 2019). A "Matthew *Mechanism*" then, is a theoretical process that gives rise to the observed Matthew Effect. According to Bask & Bask (2015), by definition a Matthew Mechanism must 1) be a dynamic process that results in inequality of outcomes; 2) magnify small initial (dis)advantages such that even a slightly

85th Annual Meeting of the Association for Information Science & Technology | Oct. 29 – Nov. 1, 2022 | Pittsburgh, PA.

favorable relative initial position produces further relative gains; and 3) vary across the affected population. Prior work has eschewed identifying mechanisms, focusing instead on awardee attributes while largely ignoring collaborators' statuses as predictors of R01 (or equivalent) success and predicting success from variables' values at single points in time without regard to the full history of value changes leading up to those observed values. Lin (1999) defines social capital as "*investment in social relations by individuals through which they gain access to embedded resources to enhance expected returns of instrumental or expressive actions*."

We represent MK awardees' social capital relationships as a graph/network where nodes are MK awardees and their coauthors, and their relationships (e.g., co-authorship) are represented by edges (links) to indicate shared publications. Thus, each MK awardee has an ego-centric co-authorship network in which the resources (social capital) of those coauthors vary by person and time, and consist of reputation, status, wealth, and power. These attributes represent potential resources embedded in the professional networks of MK awardees that can be leveraged in their quest for an R01. These ego-networks also represent "brokerage social capital" (Burt, 2005) that may help awardees identify and capitalize on grant-funding opportunities or introduce them to potential future collaborators. Importantly, these forms of network social capital represent "relational assets" not evaluable solely by reference to MK awardees' own attributes. As such, the hypothesized process of differential social capital accumulation meets the criteria of an effective Matthew Mechanism.

Our graph/network, by design, contains the ego-centric scholarly networks of MK awardees which is necessary for selecting our predictors for our time-to-event statistical models. This is because the Matthew Mechanism is a between-persons process—small initial differences in social capital between MK awardees gives rise to differences in the process of social capital accumulation over the transition period, ultimately resulting in large advantages for the initially advantaged group with respect to likelihood of R01 funding. This is why at the core of our study we focus on ego-centric co-authorship networks as the sources of social capital in addition to individual and institutional attributes.

STATISTICAL MODELING

Statistical models have historically been employed in a great deal of scientific research. The application of probability theory to samples (ideally random) of data, through such statistical models, provides the ability to describe samples of data and their distributions, as well as providing the ability to identify associations between variables in a sample of data, documenting relationships between variables. Statistical models have been widely employed because of their perceived ability to bring rigor, and ideally, reproducibility, to the evaluation of theoretical expectations; these being most often, though not always, expressed as hypotheses. Classical multiple-regression models have long been used to analyze both cross-sectional and longitudinal samples of data where the "outcomes" of interest are continuous (or at least interval) in nature, and generally complete "follow-up" is available on all observations in the sample. The assumption of linearity in the associations between "predictor" variables and outcome variables apply to such models. Multiple logistic-regression models have been applied in the same way, but where outcomes of interest are dichotomous or polytomous. Logistic-regression models assume a log-linear relationship between predictor and outcome variables. One limitation of both of these types of classical regression is that they are not designed to handle samples of data where there is "right-censoring"—i.e. the outcomes of interest are not known for all sample observations during the period of follow-up observation. A second limitation is that such models generally ignore the role that various dimensions of "time" may play between the predictor and outcome variables. Several salient dimensions of time that may be important in understanding various phenomena include duration from some salient period or event, age of observations, and historical "period." While one or more of these dimensions can be incorporated into statistical models, classical-regression models are less suitable for assessing samples of data where the timing of the outcome itself is of central interest.

We are interested in observing the process by which biomedical scientists who have demonstrated a particular level of career success (through having obtained a traditional "training grant" from the US National Institutes of Health), do or do not succeed in proceeding to their next level of career success (obtaining a traditional "independent investigator grant" from the NIH). This is clearly a phenomenon taking place over time (duration from training grant award), that may be subject to both historical effects (historical period), as well as the biological age of the scientist at time of training grant award. It is also the case that we observe different lengths of follow-up time for individual scientists, and for some of them, we do not observe the entire duration of follow-up that may be pertinent, so data are right-censored. Moreover, our theoretical expectations lead us to anticipate that there may be time-dependence in the associations between predictors of interest and the outcome of interest. We are interested not only in *whether* the outcome of interest has occurred, but the *rate* at which it occurs. For such outcomes, time-to-event statistical modeling is more appropriate than classical regression models. Thus, we propose to employ survival regression models, which have the benefits of appropriately handling right-censored data, while specifically modeling the instantaneous rate at which the outcome of interest occurs, and the capacity for assessing non-linear associations between predictors and outcomes. Such models have the added benefits of being able to incorporate time-varying

predictor variables. The benefit of this in our situation is that many of the putative predictor variables of interest are not static, or fixed at the start of observation, but have values that evolve over the follow-up period. Examples include the number of publications a scientist has, who their co-authors and collaborators are, how co-authors are connected to one another, the institutions with which they are affiliated, etc. These features all represent forms of social capital that may influence the likelihood of obtaining an independent investigator grant. Our theoretical expectations are that the likelihood of obtaining an independent investigator grant is a function of the accumulation over time of advantages and disadvantages, associated with the accumulation of such social capital over the early course of scientists' careers.

While such time-to-event statistical models provide an appropriate way to represent and test our theoretical expectations in a rigorous way, applying the logic of probability and the ability to falsify our hypotheses, they are not without their limitations. In particular, such models are limited by our current understanding of the phenomenon of interest, and how broad and complex a range of predictor variables can be assessed directly.

This latter shortcoming could be addressed by developing latent-variable measures from directly observable measures, and using the resulting latent variable constructs as predictors in our statistical models.

The contributions of the statistical modeling approach to this interdisciplinary research venture include the provision of an explainable model, grounded in theoretical expectations, that can be contrasted to the DL model. In addition, the time-to-event model may be improved by using latent variables identified by data-driven methods as model inputs. Moreover, the statistical model can serve as one indicator of the extent to which the current state of social capital theory explains R01 success.

AI: DEEP LEARNING ON HETEROGENEOUS, TEMPORALLY-DYNAMIC GRAPHS

Contextually-Rich, Longitudinal Bibliographic Data is increasingly being deployed in various AI applications, especially for decision support and notably in domains where there is heightened sensitivity due to the protected nature of the data and due to the fact that the decisions may have serious consequences, e.g., in healthcare and legal domains. Most of these approaches emanated from computer science and have lacked involvement from other disciplines, e.g., information science and digital library communities. While powerful for prediction, these methods have not yet been fully exploited nor widely adopted due to apprehension regarding trustworthiness, potential for bias, and lack of interpretability/explainability of results. Dr. Liu will give an overview of some AI methods he is using to understand temporal relationships and their context within heterogeneous bibliographic and NIH award data. These include hierarchical community detection and heterogeneous graph mining for global, local, and hyper-local graphical feature extraction. His presentation will emphasize issues of trustworthiness in AI, model explainability and AI-output interpretability. A challenge we faced was in trying to understand relationships among multiple types of entities, e.g., scholars, research topics, publications, institutional affiliations, funders etc. Moreover, those relationships may be multiplex and evolve over time. For example, two authors may share one or more relationships as: co-authors on a publication; members of an academic institution; award recipients from a specific funder or award. Analysis of this type of complexity is challenging, especially for large data sets. To address this, we use an approach to representing the data in graph format whereby data are indexed in the same data hub, and innovative AI models applied to address complexity, heterogeneity and scalability problems. This approach also helps scholars from other areas to easily access, understand and consume the AI deliverables and sets the data and analysis up to be compatible with aforementioned approaches in this panel (i.e., theory-based, statistical). Training AI models to perceive data and model trustworthiness, given a downstream task, can be of critical importance. Yet, task-oriented data quality detection and enhancement have been much less studied by the AI research community and is often considered as an afterthought. Dr. Liu will demonstrate how his approach to explainable AI and trustworthy AI enables important collaborations between AI and other research communities, and in particular he will show how data-driven (AI/DL/ML) models for identifying latent features can identify predictors for statistical models.

INTERDISCIPLINARY INTEGRATION: EXPLAINABLE AI TO ADDRESS TEMPORALLY COMPLEX PHENOMENA

Dynamically complex relationships are frequently present in the research questions we aim to address—for example human health, global climate, and scholarly success can all be conceptualized a states that evolve over time. And yet, such problems are often analyzed according to oversimplified conceptual models that fail to adequately incorporate the full contextual dependence on, and complexity of - time. As humans, researchers are prone to the "Flatland Fallacy" (relying on overly simplistic models of reality that hew to our perceptions, mistaking them for the complete picture of the phenomena, which is far more complex (Jolly & Chang, 2018). Researchers are also biased in favor of model parsimony, which we commonly over-subscribe to—leaving out the very complexity we seek to understand. Tragically, the error is quite common, and with respect to time varying phenomena, variables of interest are measured once or at a few select points in time and analyzed under the assumption that they operate as linear

processes (if they are animated at all), obscuring time's true identity as a continuous variable capable of marking a myriad of dynamically changing contexts. Heuristics are human artifices which simplify reality, enabling us to make quick accurate decisions (Kahneman, 2011)…at least under ordinary circumstances. When employed under conditions of temporal complexity, such when trying to grapple with heterogeneous dynamic processes, oversimplification can result in bad decisions which can trigger unintended consequences of disastrous proportions (Sterman, 2006). Therefore, it is crucial that we develop appropriate designs and methods to appropriately capture, analyze and interpret time-dependent, contextualized relationships.

Continuing with the exemplar case-study of trying to understand how social capital processes might underlie differences in scholarly success, Dr. Mabry will describe how interdisciplinary research designs utilizing elements of theory-based, statistical and data-driven (AI) methods can leverage the strengths of each to launch us further in our nascent development of tools for understanding dynamic processes. For example, she will describe how this team is using data-driven methods as a benchmark to evaluate our theory while using theory to explain and interpret results from our predictive models. She will recount some of the challenges and rewards of interdisciplinary work encountered in this project and explain how some of these challenges have been handled. She will then moderate the discussion to follow the panel.

CONCLUSION

The proposed panel will give the audience an opportunity to engage the presenters in a discussion of challenges faced when conducting research on dynamically complex heterogeneous big data, and this team's unique interdisciplinary approach to research design and analyses to overcome them. All presentations will be illustrated using the team's NSF-funded project exploring how social capital may accumulate over career trajectories and lead to scholarly success, to facilitate discussion of sophisticated and diverse disciplinary approaches to understanding temporally complex big data. The discussion will be designed to engage the audience in a broader discussion of how to improve research design and analytics for dynamically complex research questions and to stimulate thinking on how the methodologies of theory-based, hypothesis-testing and data driven approaches can be combined to more effectively turn data into knowledge.

ACKNOWLEDGMENTS

We thank Holly Woodward and Alyse Haven, Project Managers, HealthPartners Institute, for their assistance in putting this proposal together for submission and her faithful dedication and active involvement in managing our research project that served as a case-study for this panel. We also gratefully acknowledge support from NSF, Award # 2122232 - SCISIPBIO: *Constructing Heterogeneous Scholarly Graphs to Examine Social Capital During Mentored K Awardees Transition to Research Independence: Explicating a Matthew Mechanism*.

REFERENCES

Bask, M., & Bask, M. (2015). Cumulative (Dis)Advantage and the Matthew Effect in Life-Course Analysis. *PLOS ONE*, *10*(11), e0142447. doi: 10.1371/journal.pone.0142447

Burt, R. (2005). *Brokerage and Closure: An Introduction to Social Capital*. New York: Oxford University Press.

Jolly, E., & Chang, L. (2018). The Flatland Fallacy: Moving Beyond Low–Dimensional Thinking. *Topics In Cognitive Science*, *11*(2), 433-454. doi: 10.1111/tops.12404

Kahneman, D. (2011). *Thinking, Fast and Slow* (1st ed.). New York: Farrar, Straus and Giroux.

Katz, Y., & Matter, U. (2019). Metrics of Inequality: The Concentration of Resources in the U.S. Biomedical Elite. *Science As Culture*, *29*(4), 475-502. doi: 10.1080/09505431.2019.1694882

Jiang Z, Gao Z, Lan J, Yang H, Lu Y, Liu X. Task-Oriented Genetic Activation for Large-Scale Complex Heterogeneous Graph Embedding. Proceedings of The Web Conference 2020: Association for Computing Machinery; 2020:1581–1591.

Jiang Z, Yin Y, Gao L, Lu Y, Liu X. Cross-language Citation Recommendation via Hierarchical Representation Learning on Heterogeneous Graph. The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval; 2018; Ann Arbor, MI USA.

Lin, N. (1999). Building a network theory of social capital. *CONNECTIONS*, *22*(1), 28-51.

Merton, R. (1968). The Matthew Effect in Science. *Science*, *159*(3810), 56-63. doi: 10.1126/science.159.3810.56

Merton, R. (1988). The Matthew Effect in Science, II: Cumulative Advantage and the Symbolism of Intellectual Property. *Isis*, *79*(4), 606-623. doi: 10.1086/354848

Sterman, J. (2006). Learning from Evidence in a Complex World. *American Journal of Public Health*, *96*(3), 505-514. doi: 10.2105/ajph.2005.066043