**Poster Submission**

## Using Digital Trace Data to Examine the Social Dynamics of Scientific Teams

Digital trace data is a by-product of both human and computer system activity. This digital trace data can be harvested and used to identify patterns of activity – in the consumer domain, we often see this manifested as social networking recommender systems or targeted advertisements. However, it is our supposition that digital trace data can also be used to better understand the social dynamics of teams; and in our case, scientific research teams.[1] Our research utilizes a combination of digital trace data and semi-structured text data sources to study the activities of scientific teams to build an understanding of the social dynamics of scientific work using an analytical and computational approach. This poster illustrates our approach and early findings of the research.

Today's scientific teams often utilize a computational capability in addition to laboratory or field study to conduct their work. The scientific teams we are most interested in understanding are those that are composed of a heterogeneous mix of experts working together to leverage each other's skills and knowledge to accelerate time-to-discovery of the phenomena of interest. Where, 'discovery' in our context means gaining new insight or understanding and central to the work is the analysis of massive and/or complex data.[2] For example, the mix of pharmacological, oncological, computer science, and mathematical skills to develop machine learning and natural language processing algorithms to aid in the discovery of treatments for diseases such as cancer, Alzheimer's, or Amyotrophic Lateral Sclerosis to enhance and focus wet lab research and development.[3] The insights gained into the social dynamics of these scientific teams (e.g., collaboration, operation, information sharing) are used to improve two dimensions of their project work: (a) processes affecting team operations, such as information dissemination practices; and (2) cloud/computer infrastructure improvements, such as supported software configurations.

The approach to develop and utilize this method of examination started with the selection of appropriate data that would afford examining interaction and activity in a project and a corresponding method of analysis for each type of data. Our goal was to identify disparate data that could be compared over the same time period to reveal insights and patterns not visible using a single form of data. We are interested in understanding both common and unique patterns of social dynamics within and between projects. Our initial examination used a corpus of email data that was analyzed using network analysis and content analysis. Network analysis provided insight into communities and interactions among the project participants. Content

---

[1] Pentland, A. (2014). Social physics: How good ideas spread—The lessons from a new science. New York: Penguin.

[2] Haas, L., Cefkin, M., Kieliszewski, C., Plouffe, W., & Roth, M. (2014). The IBM Research Accelerated Discovery Lab. *ACM SIGMOD Record*, *43*(2), 41-48

[3] http://www.research.ibm.com/client-programs/accelerated-discovery-lab/index.shtml

analysis using Leximancer identified natural language concepts and themes and emerging and fading topics of conversation.

Early results are promising for this approach of combining heterogeneous sources of data, including "thick data"[4] to identify patterns of science team activity. We have been able to identify temporal evolution of work activities, primary actors and/or stakeholders during different phases of work, and conversational aspects as the work evolves, matures, and then ends. For example, our results illustrate the arrival and departure of project team members (i.e., managers, researchers, and executives) along with conversational variations over the project lifecycle that were both expected and unexpected.

We have had success with two data sources, however we think this only provides a partial picture of activity. Building on our current findings, (a) the inclusion of additional data sources (e.g., such as system logs[5], meeting transcripts, and project artifacts) are a high priority to deepen understanding and establish repeatability, and (b) the identification and development of appropriate metrics to gauge the pace of science team activity. Once the analysis of the data is complete, our intention is to then use the results to develop an agent-based model[6] to perform what-if scenarios to improve team dynamics and system effectiveness with the desired result of recognizing, enabling and accelerating innovation, and fostering scientific discovery.

---

[4] For example, http://ethnographymatters.net/blog/2013/05/13/big-data-needs-thick-data/

[5] Dumais, S., Jeffries, R., Russell, D. M., Tang, D., & Teevan, J. (2014). Understanding user behavior through log data and analysis. In *Ways of Knowing in HCI* (pp. 349-372). Springer New York.

[6] Anya, O., Moore, B., Kieliszewski, C., Maglio, P., Anderson, L. (2015). Understanding the practice of discovery in enterprise big data science: An agent-based approach. In proceedings of 6th International Conference on Applied Human Factors and Ergonomics (AHFE 2015) and the Affiliated Conferences, (pp. 4389 – 4396). Elsevier Science Direct.