Winter Simulation Conference

Team Russell - Text Analytics Project

Arindam Bhattacharya, Collin Rooney, Rose Thomas, Julia Greenberger, Kevin Zhai

---

I. Executive Summary

We conducted data analysis on papers across 48 years of the Winter Simulation Conference, the foremost academic conference focused on computer simulation research. Through our exploratory data analysis and topic modeling, we were able to uncover clear trends in specific topics appearing each year, ranging from the evolution of simulation software and tools to appearance of social and sustainability-oriented topics in recent years. We believe these findings provide valuable insights into the past, present, and future of the WSC.

II. Problem Statement

Dr. Barry Nelson, a professor in the Industrial Engineering department at Northwestern University and a leading figure in the area of simulation, approached our team with a business problem. Dr. Nelson has very close ties to the conference, as he has participated in the conference for more than 20+ consecutive years, contributing numerous research papers and presentations about his research. He was selected as the keynote speaker for the 50th anniversary of the WSC in 2017. In preparation for such a noteworthy occasion, Dr. Nelson thought that it would be worthwhile to analyze the history and evolution of the conference since its inception in 1968.

Our primary tasks involved collecting and analyzing high-level information (titles, authors, etc.) associated with the archived papers stored for every year of the conference. Through our analysis, we aimed to assist Dr. Nelson with his keynote speech by finding interesting historical trends in topics, anomalies that happened in certain years, or other insights that proved to be enlightening and/or entertaining.

III. Business Value

We hope that our basic visualizations can provide an idea of which authors at the conference have been the most influential. It is our intent that our mapping of the institutions will provide a compelling illustration of the growing diversity of attendees. The topic modeling portion of this project was completed with the goal of providing Dr. Nelson and his audience historical information about software trends. We hope that these findings will give Dr. Nelson more options and information to work with when planning his presentation next year.

IV. Technical Methods

Data Gathering (Web Scraping)

In order to gather the necessary data and transform it into an analysis-friendly format, we used R and the package "rvest" to write several scripts that scraped relevant data from the Winter Simulation Archive website for the years 1997-2015. The layouts and underlying HTML structure of these webpages varied greatly from year to year, preventing us from creating a single general script that could obtain all of the requisite data. Instead, we wrote separate scripts for groups of years, with webpages for each group having sufficiently similar layouts to render a single script feasible.

After extracting the raw data from the website, we performed considerable post-processing. The authors associated with each paper and their corresponding affiliations were extracted as a single string, so we wrote a script to parse and tokenize these strings into individual authors and affiliations. We identified several authors and affiliations of interest, and manually cleaned up the data associated with them (i.e. consolidating records with slightly different spellings of an author's name). In addition, abstracts for each paper were only available from 2011 onwards, so we manually entered abstracts for the remaining years into the data set. Finally, we normalized our data set and created a Postgres database.

EDA and Basic Visualizations

After obtaining, cleaning, and formatting our data, we performed basic exploratory data analysis (EDA) of authors and affiliations, hoping to uncover points of interest in the process. We began by constructing a list of highest contributing authors in which contribution was measured as number of papers from 1997-2015. We then visualized the results in a bar chart (Figure 1). Doing so

helped us identify several “titans” of modern simulation, and our results were in line with expectations; Dr. Nelson commented that he was close colleagues with nearly all of the authors identified in our visualizations.

We then shifted our focus to institutions, identifying the highest contributing American and international institutions, using the same metric as before. We mapped these results, producing the visuals in Figures 2 and 3. Among American institutions, we found that Georgia Tech was far and away the most “productive” university - a result that was again supported by Dr. Nelson’s description of GT’s simulation faculty and resources. Internationally, we found 2 major clusters of elite institutions - Europe and, somewhat surprisingly, Singapore, which houses 2 of the top 20 institutions (as defined by our previous metric).

Mapping Institutions

Once all institutions were scraped from the archive, we used a R function called “geocode” to gather the longitude and latitude values of these organizations. This function required a character input that would then be matched to a set of coordinates by Google Maps. If the character string could not be matched, the function output a value of “NA.” Many institutions were not found, and as a result, many NA’s were recorded. The non-missing values were then analyzed to see if they were contained within the United States.

We constructed a box of latitudinal and longitudinal points to represent the boundaries of the United States. If the individual institution’s coordinates were contained within this box, the affiliation was labeled as a domestic affiliation. If the coordinates were outside the box in any direction, the affiliation was labeled as foreign. The labeling of these affiliations were completed in Excel by using a few if statements and equations. The Excel table of affiliations, years, coordinates, and labels was then loaded into a Tableau project.

Using Tableau’s mapping feature, the team was able to explore the affiliations visually and find interesting insights that could be useful to Dr. Nelson. Figure 4 shows one of the first findings on foreign institutions. The number of affiliations outside the United States seemed to greatly increase between the years of 1997 and 2015. We did not know how much of this difference could be explained by varying attendance or NA values, so we subsequently analyzed the percentage of foreign affiliations.

To complete this task, we calculated a ratio in Excel of institutions found to be outside the United States against total number of conference attendees for each year. These yearly percentages

and the least squares regression line are plotted in Figure 5. All variables and coefficients in this simple linear regression model were found to be significant. From this analysis, we were able to provide Dr. Nelson with an interesting and significant finding that he can share at his upcoming conference.

Topic Modeling

After our initial EDA, which uncovered several high-level points of interest with respect to the history of the conference, we took a more granular look at how the conference has evolved over time through topic modeling.

Although abstracts contained the highest level of detail among our available data fields and would have been a natural choice for topic modeling, they were only available from 1997 onwards and our goal was to find trends across the entire timespan from 1968-2015. As a result, we performed topic modeling on paper titles, which still contained sufficiently detailed information to be meaningful. We aggregated all titles for a given year into a single document, and our final input into our topic model was a document-term matrix with 48 documents, with each document representing one year.

Prior to running LDA on our document-term matrix, we performed several pre-processing steps. We began by removing stopwords and punctuation, as well as tokenizing the data into unigrams (single words). Next, we set hard thresholds on token frequency across documents - in order to remain in our final DTM, a word had to appear in at least 2 different years and in no more than 30. This removed many words with little information content (e.g. “simulation”, “paper”, etc.). In order to further pare down the number of words in consideration, we calculated tf-idf scores for each token and removed those with scores below the median.

Finally, we ran VEM (variational estimation maximization) LDA on our DTM, focusing on the top 30 words associated with each topic and experimenting with various numbers of topics. Although the typical idea is to use a much smaller number of topics than documents, we found that this approach led to very generic, similar topics that did not provide any meaningful insights into how topics changed over time. Consequently, we tried a different approach set the number of topics equal to the number of years.

Although unorthodox, we found the results to be much more meaningful and interpretable, with clear differences between years and historical trends coming to light through careful analysis of the resulting topics. For example, an interesting anomaly we found through topic modeling was the

presence of “criminal justice” in a single year, 1977. It turned out that there was an entire session dedicated to the topic in just that one year, and that it has otherwise remained a relatively unexplored application of simulation (judging by the contents of the conference). Given the emergence of criminal risk assessment scores in the judicial system in recent years, it is entirely possible that this topic may make a comeback at the conference, perhaps providing a glimpse of what the future of the conference holds.

Software Evolution

The topic modeling output assigned the corpus of conference paper titles from each year to one topic. The topic modeling output was a 30x48 matrix (top 30 words for each year). We used this output to extract trends in research areas and simulation software, as well as identify any anomalies. Upon closer examination of the words in each topic, we found several software languages and tools associated with simulation.

To visualize how these softwares and tools overlap and evolve, we extracted the names of simulation softwares from topic modeling. We then plotted with a red square each year there was a mention of this software in any paper title at the conference. The black bars span the first through last mention. From Figure 6, we see the most widely mentioned software since 1968 is GPSS, which has decreased in popularity since 2000, likely due to the emergence of new programs such as Java. We believe it would be interesting to revisit this analysis in a few years to see if open source simulation softwares gain more relevance in conference paper titles.

Research Area Trends

To visualize how research areas in the simulation community have evolved, we created an interactive word cloud as seen in Figure 7. Using terms from each year’s topics, the word cloud shows the most prominent topics. Our Tableau visual allows the user to hover over a word to display the years in which the term appears in the topic. We found an interesting trend in papers researching “energy.” A prominent topic in early conference years, energy did not appear in our topic analysis for nearly three decades, until a resurgence of interest began in 2009.

V. Appendix

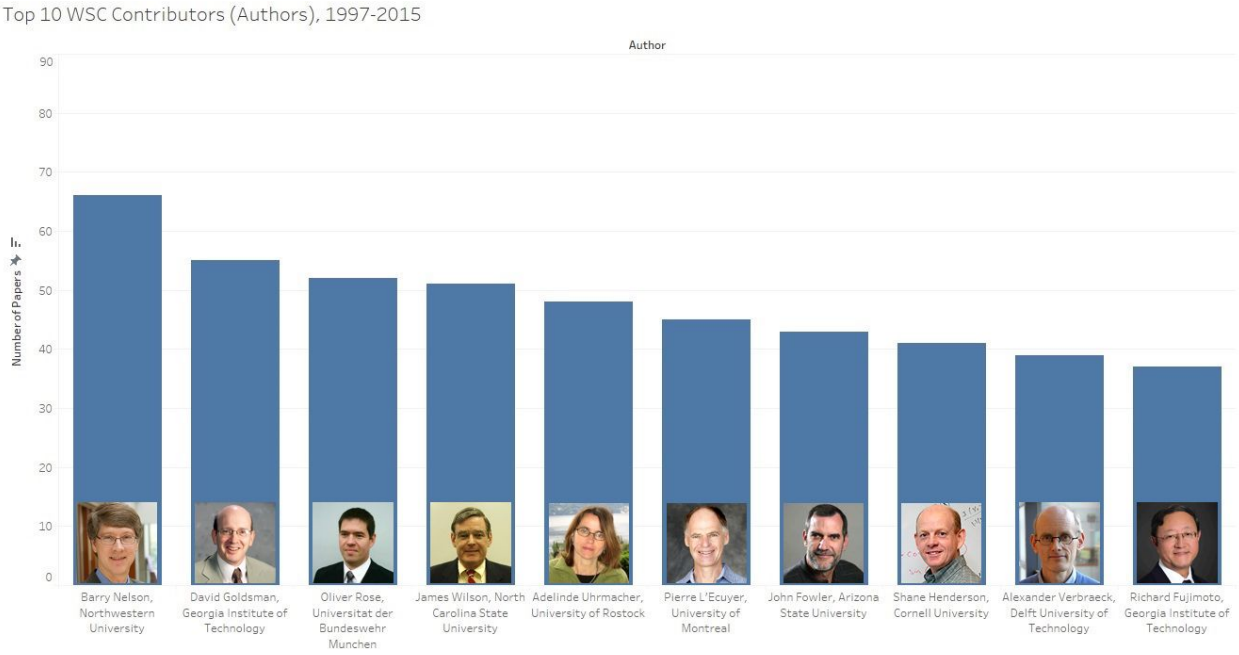

Figure 1: Top WSC Contributors (Authors), 1997-2015

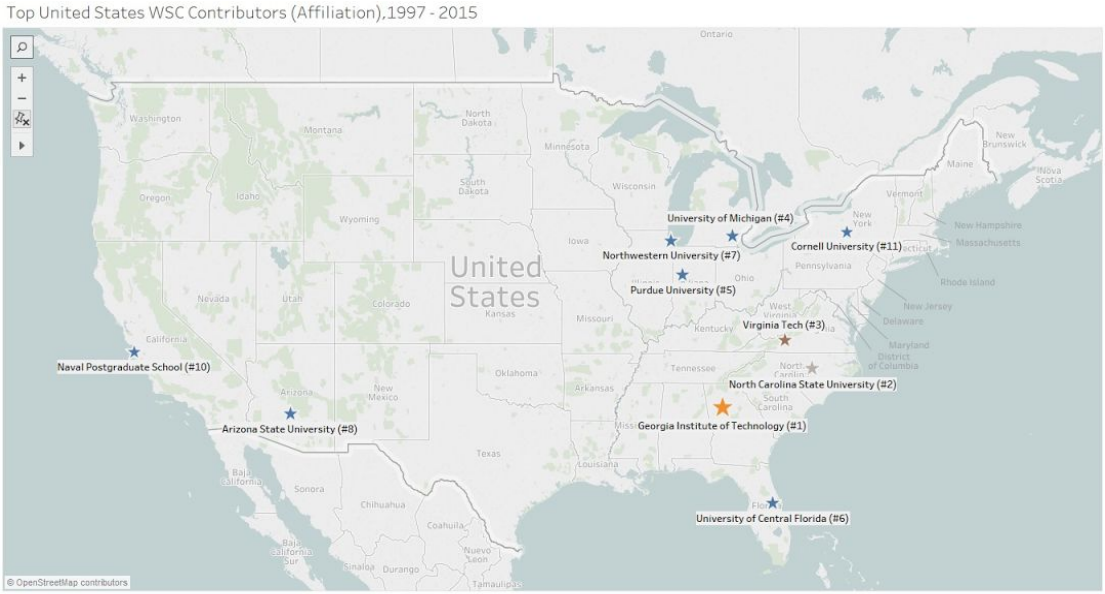

Figure 2: Top WSC Contributors (U.S. Institutions), 1997-2015

Figure 3: Top WSC Contributors (International Institutions), 1997-2015

Figure 4: Foreign Affiliations in 1997 and 2015

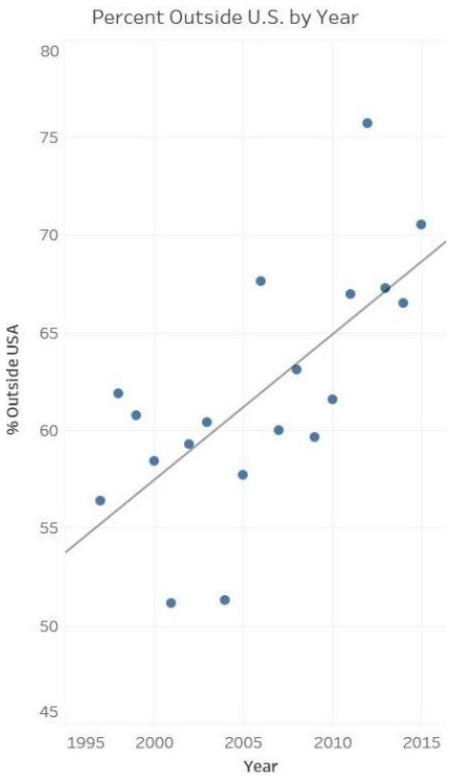

Figure 5: Percent of Affiliations Outside the U.S. by Year

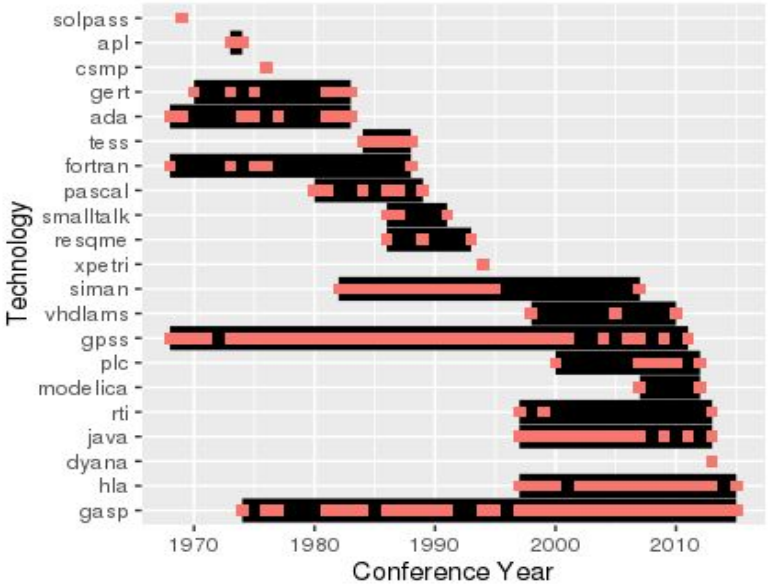

Figure 6: Appearance of Simulation Languages by Year

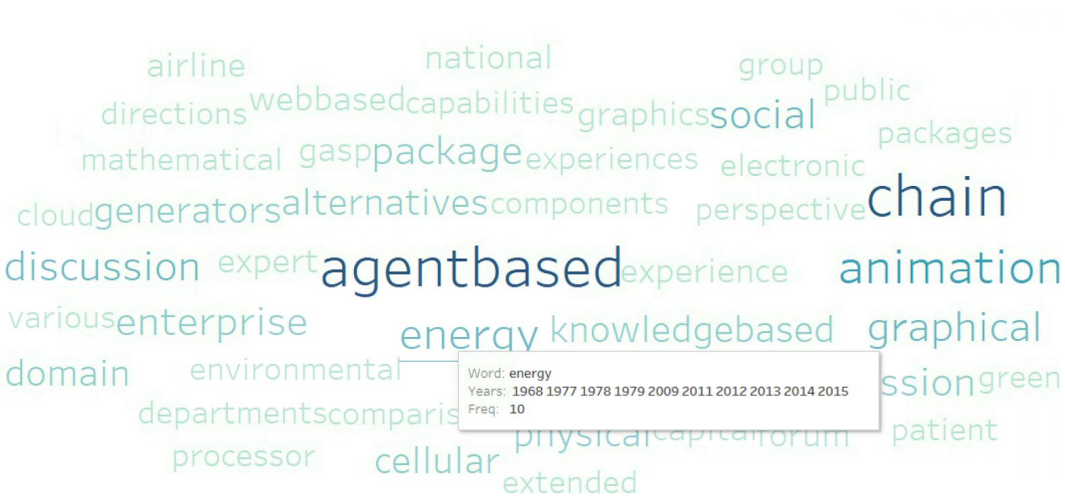

Figure 7: Word Cloud of Topic Modeling