Moses Boudourides
Research Statement

My research interests and work focus on social network analysis, network science, computational social science and digital humanities. The list below displays a number the research topics of my recent (during the last 5 years) and ongoing work.

1. *Homophily in Patients' Geographic Areas, Ages etc. in COVID-19 Contact Networks*:

The dataset that we are processing and analyzing is a multi-level dataset of COVID-19 patients in Bucharest, Romania, during 2020-21 (two years). At the node-level, one gets a directed graph among patients and their contacts (i.e., people with whom the former have interacted before their first COVID-19 symptoms). At the nodal attribute level, there are data on patients' and contacts' geo-locations (at the street level), their age and a few more socio-demographic characteristics. Our aim is to aggregate geo-locations (age etc.) as the nodes of a reduced directed graph and to study the aggregated patterns of edges as indicators of varying occurrences of homophily (or assortativity) or mixing (or disassortativity) in different time periods. The structural part of our analysis is based on the temporal evolution of the mechanism of triadic completion (or closure) in the directed graph of patients and contacts. Actually, the connectivity patterns of triadically open paths of length 2 constitute the structural mechanism for the epidemiological propagation. Thus, analyzing the temporal labelled graph of patients and contacts and detecting the patterns of how triadically open 2-paths are temporarily aggregated and distributed across geographical areas (and other groups of the sociodemographic variables) helps us understand how the virus is spreading among different neighborhoods and areas in cities.

2. *Preference Aggregation through Degree Centrality Indices*:

Suppose each one of *N* voters issues a list of preferences on *M* candidates. Without any loss of generality, we assume that these preferences are linear and that they can be partial (i.e., not all candidates might be ranked by all voters). Then using the fact that every voter's preference is composed of a number of ordered binary preferences (as those ordered pairs of candidates included in voters' relational preferences), a multiple (weighted) digraph is generated, every edge of which is created by "summing" (aggregating) all voters' binary preferences on the candidates who are end points of the edge (thus, with *N* being the maximum possible weight on an edge). Let us call this graph "graph of voters' binary preferences" or, in short, binary-preferences graph. Having the latter graph, we can correlate the counted preferences of all voters (computed by various algorithms of voting aggregation in social choice theory, like Borda, Carroll, Black, Kemeny and others) with the ranking of various centrality indices of nodes-candidates of the weighted digraph of voters' binary preferences. In this way, we are expecting to determine which centrality index is optimally approximating which voting method of preference aggregation. (https://medium.com/@mosabou/cumulative-rank-aggregation-of-a-family-of-network-centrality-indices-e625a76bf7e4)

3. *Leveraging the Study of Student Mobilization with Machine Learning*:

We are using the powerful machine-learning framework of word embeddings in order to analyze the archive of resources for the 2020 Haverford College Student Strike. These resources include documents created by Haverford student strike organizers and participants, inactive students, and members of the Haverford faculty, staff, and senior administration. The collected pieces of text of this archive include email correspondence, departmental statements, news articles, social media posts, transcripts of student town halls, negotiation meetings, and teach-ins, organizational materials pertaining to food distribution and mutual aid and much more. This corpus is publicly available in a dedicated archive maintained by the TriCollege Libraries at https://digitalcollections.tricolib.brynmawr.edu/collections/haverford-2020-student-strike-collection. Through a word embedding model, which is a commonly used tool in natural language processing (NLP) and machine learning, we are developing a computational linguistic framework to measure, quantify, and compare the formation and the time evolution of the debate and the narratives uttered by students and other actors in the events of the 2020 Haverford College Student Strike. The aim is to identify and situate those linguistic traces of social categories, with regards to individual and collective action and social institutions, which signal and fuse at the intersection of academic, educational, cultural, economic, and social spheres, as they are discursively constituted and differentiated by race, gender, age disparity, academic and disciplinary authority and political power. Furthermore, deriving and analyzing the network of co-occurrences of these word embeddings, our goal is to apply techniques of social network analysis and network science to gain a better understanding of the intersectional generative mechanisms framing the constitution and the time evolution of the 2020 student mobilization at Haverford College.

4. *Networks of Wikipedia hyperlinks*:

Starting from a small number of Wikipedia pages centered around a common theme (examples of case studies that we have already investigated: Medieval Literature, Futurism, Poststructuralism, Climate Change etc.), we are collecting (using Python) all other Wikipedia pages hyperlinked by the former. After detecting the existing hyperlinks among all collected Wikipedia pages, we are studying the resulting directed graph of hyperlinks as a multi-ego-centric network (in which the initially selected starting Wikipedia pages are the egos and their hyperlinks their alters). Furthermore, we are detecting certain important egocentric subnetworks inside the overall graph centered on various particular Wikipedia pages, which are nodes of the overall graph of hyperlinks and the content of these pages is of salient or focal interest. Moreover, we are determining paths of access and (geodesic) graph-distances and cycles (or recurrent motifs) among the pages. Finally, we are studying the assortativity/homophily of the overall network of hyperlinked pages depending on the origin of a page from hyperlinks, which are descending from one or more than one of the starting Wikipedia pages. (https://medium.com/@mosabou/the-citation-network-among-wikipedia-pages-on-dynamical-systems-and-mechanics-7b41e1ae728)

5. *Dominating Intersection Subgraphs from Wikipedia Pages Associated with Christian Nationalism*

To identify connections and power relationships among Christian Nationalist and adjacent institutions and individuals, we have analyzed the connections between over 6000 Wikipedia pages associated with Christian Nationalist subgroups, institutions, and individuals. The pages were collected from the Wikipedia category indices for subgroups and issues related to Christian Nationalism. Our network analysis was motivated by the scope of graph domination theory. Denoting by $N(u)$ the neighbors of page u in the Wikipedia hyperlink network, we considered a coupling coefficient between any two pages u and v defined as: $c(u,v) = |N(u) \cap N(v)| / \min(|N(u)|,|N(v)|)$ (where $|.|$ denotes number of elements). Since $N(u) \subseteq N(v)$ or $N(v) \subseteq N(u)$ if and only if $c(u,v) = 1$, one might classify three types or relations of dominance among any two pages u and v: (i) u "dominates" v whenever $c(u,v) = 1$, (ii) u "intersects" v whenever $0 < c(u,v) < 1$ and (iii) u "detaches" v whenever $c(u,v) = 0$. Furthermore, since for every page u there exists a “maximal” dominating page v containing u and not contained in (the neighborhood of) any other page, one might consider the induced subgraph of “maximal” dominating pages as an intersection graph. In this sense, the resulting “maximal dominating intersection graph” preserves the structure of power and dominance among pages. Thus, our aim was to study the pattern of major pages in the dominating intersection graph in order to interpret how Wikipedia tends to represent the power structure, the partnership and the polarization among various individuals and Christian Nationalist organizations. (https://github.com/mboudour/var/blob/master/CNAG.pdf)

6. Networks of affect in the *COVID19positive* subreddit posts*:*

Our aim was to explore from both data- and network-analytic point of view the dynamics of affect developing in the posts of the *COVID19positive* subreddit. Our dataset included all the threads of discussions on this subreddit, during 181 days in 2020, from the commencement of the subreddit on March 14 until September 11, 2020. In this period, the *COVID19positive* subreddit has hosted 12624 posts by 1406 redditors, who have contributed 167306 comments. Using standard techniques of NLP of POS tagging, from the corpora of posts of these discussions, we were able to extract a big dataset of employed in the text verbs in stemmed form. Based on standard linguistic methodologies, the extracted verbs were partitioned into four categories: action, doxastic, emotive and sensory verbs. Our research hypothesis was that the strongly affective features of the discourse developed in social media around the current COVID19 pandemic entail an increasing use of verbs in all these categories depending on the context and the dramatic character of the discussions. Moreover, verbs appear to co-occur in sentences, the sentimental (analytic) score of which tended to increase when the pandemic situation happened to escalate. For this reason, we were studying temporal networks of sententially co-occurring verbs in the *COVID19positive* subreddit in order to trace the discursive ways in which feelings, concerns, opinions and emotions were embroiled, discussed and were unfolding the experiential narrative of the pandemic

inside Reddit.
(https://drive.google.com/file/d/1OYrF1ef3YcYji1seWL9x_qWoe7EzYGZe/view?usp=sharing)

7. *Data and network analysis of Dictionaries or Lexicons*:

We were working on text data from existing Dictionaries or Lexicons (examples: Niall Lucy's "A Dictionary of Postmodernism", Voltaire's "Dictionnaire philosophique" ["Philosophical Dictionary"], Raymond Williams' "Keywords: A Vocabulary of Culture and Society" etc.). Typically, such dictionaries contain descriptions and discussions on a number of principal terms (concepts or ideas). These terms are lexical items and in the content of presentation of each of them in the Dictionary, other lexical items (terms) are mentioned and referred to too. Moreover, the contents of each lexical item may include a number of references to various published works (in the literature of the field of the Dictionary). In other words, we were deriving a directed network among Dictionary terms and (possibly) a two-mode network among such terms and cited authors (in possibly existing scholar references). In this way, we were studying the structure of the overall network of terms of the Dictionary and its assortativity/homophily according to citations to particular cited works. Finally, using the text-analytical methodology of Topic Modeling, we were identifying (through tf-idf) a vocabulary of important words in the corpora of all Dictionary items and studying the resulting word-net from co-occurrences of these words inside the contents of all lexical items. (https://medium.com/@mosabou/the-graph-of-raymond-williams-keywords-de7bb0e0a9f8, https://medium.com/@mosabou/jumbling-up-two-assemblages-the-networks-of-the-deleuze-and-meillassoux-dictionaries-e2e224121ab9,
https://github.com/mboudour/var/blob/master/Boudourides_WikipediaHyperlinkGraphs.ipynb,
https://github.com/mboudour/var/blob/master/LRHEM-v0.4.pdf)

8. *Networks of bibliometric data*:

From the Web of Science (WoS) or Scopus, we were retrieving bibliographic datasets on particular fields or areas (examples: Humanities, Liberal Arts, Interactive Media, Feminism etc.), which have been published in a rather large range of years. After presenting a description of the basic statistics of these data, we were focusing on the timeseries of authors, sources of publications (journals or volumes etc.) and keywords employed by both authors and the WoS or Scopus. Furthermore, we were constructing and analyzing three kinds of networks extracted from these data: (1) citation graphs, (2) co-authorship networks and (3) keyword-networks (during years or decades or in the total time period). In addition, we were studying the assortativity/homophily of the former networks with regards to selected groups of keywords s attributes at the nodal level.
(https://github.com/mboudour/var/blob/master/Boudourides_CSS_Seminar.pdf,
https://github.com/mboudour/var/blob/master/Boudourides_DH_Seminar.pdf,
https://github.com/mboudour/var/blob/master/Boudourides_RGNS2019.pdf,
https://drive.google.com/file/d/1OQ5iREHCyXuPiXzV40OpW6hg7G4l9q78/view?usp=sharing)

9. *Networks of movies or songs or books*:

From various APIs, we are collecting data on movies or songs or books (examples: IMDB, iTunes, Amazon etc.). If these data are movies, typically, we are interested in titles of movies, years of release, directors, main actors and the genres of movies. If they are songs, titles of songs, performers, singers or bands or orchestras, release dates, albums and lyrics. If they are data of book s, titles of books, authors, publishers, publication dates and possibly existing rankings of books by readers. In all these cases, after presenting a description of the basic statistics of the data, we are constructing the following general types of networks: (1) the network of artists or authors having participated in common works, (2) the network of anthologies of works with common (main) contributors, (3) the network of titles in common genres, (4) the network of directors with common (main) actors playing in their works, and (5) the network of directors with common genres of their works. Furthermore, we are examining the structural characteristics of these networks (centralities and community partitions) and their evolution in time according to their release years. (https://github.com/mboudour/var/tree/master/TextAnalysis_KeywordsCoOccurrenceNetworkCommunitiesTopicModeling, https://github.com/mboudour/var/blob/master/Boudourides_SententiallyCo-OccurrentGraphsOf2GramsInBooks.ipynb, https://github.com/mboudour/var/blob/master/Boudourides_Twitter_Seminar.pdf)

10. *Initial Boundary Value Problems of Network Processes:*

Given a graph (network) and a dynamical process on the graph, the initial boundary value problem (IBVP) of the process is to solve the equations of the process on the graph, in which case solutions vary on nodes and time. In this context, the boundary of the graph is defined as the set of those nodes, over which known values are assigned to solutions. Thus, the IBVP of a network process is to find solutions which satisfy such conditions on the boundary (Dirichlet conditions) and initially they take prescribed values on the graph. Apparently, by controlling boundary conditions on a “small” subset of nodes of the graph, one is expected to control either the equilibrium steady solutions or any possible oscillations of solutions. Among the network processes that we have applied this methodology are diffusion processes and Friedkin-Johnsen network influence processes (https://github.com/mboudour/var/blob/master/Boudourides_ExperimentsOfSocialInfluenceOnGraphs.pdf, https://github.com/mboudour/var/blob/master/Boudourides_NYUADmath2019.pdf, https://github.com/mboudour/var/blob/master/ExperimentsInFJModelOfNetworkInfluence.ipynb, https://drive.google.com/drive/folders/1_tYvNfZL5Urdrae2-Zhxff62rNE6Rn1o?usp=sharing).

11. *Google autocompleting and underlying attitudinal stereotypes*:

We were collecting data from the Google API through queries of the form "Why + X", where X is a nationality or any other identity of people. From the retrieved autocompletes by Google, we were extracting responded characterizations (for nationality X), conceived as stereotypes that Google is promoting in the autocompletion of its searches. Thus, we are getting a two-mode network of nationalities/identities/etc. vs. stereotypes and we were analyzing in the sequel. In particular, we were interested in the subnetwork of stereotypes associated to more than one nationality/identity and, thus, "bridging together" these nationalities/identities according to Google's preconceived ideas, the way they were covertly promoted to users of the autocomplete application. Furthermore, the machine-learning k-modes algorithm was used to cluster these data in a number of discrete groups. (Here is a similar work: https://medium.com/@mosabou/lost-in-translation-inside-a-multipartite-network-60650747c175)