Growing and Pruning the Republic of Letters: An Agent-Based Model to Build Letter Correspondence Networks

Buarque, Bernardo
bsbuarque@mpiwg-berlin.mpg.de
Max Planck Institute for the History of Science, Germany

Vogl, Malte
mvogl@mpiwg-berlin.mpg.de
Max Planck Institute for the History of Science, Germany

Motivation

The growing availability of historical correspondence metadata provides the digital humanities literature with valuable resources to visualize, model, and measure past socio-epistemic relationships (van den Heuvel 2015; Edelstein et al. 2017; Hotson et al. 2019; Urena-Carrion et al. 2022). However, we still have limited information regarding the inherent biases within these correspondence datasets. In other words, as Barabasi et al. (2002: 612) write about the earlier uses of scientometrics data: "for any network, before attempting to model it, we need to understand the limitations of the data collection process and*test their effect on the quantities of interest for us*." Hence, following the example set by the authors, we propose a generative model of historical letters. We introduce an Agent-Based Modelling (ABM) approach to reconstruct communication in the Republic of Letters.

Literature Review

Our model takes inspiration from two sources. First, as indicated before, we draw from earlier stochastic simulations of academic sciences. As Gilbert (1997: 91) eloquently explains, such simulations sought to reproduce the observed patterns of scientometrics "using a small number of simple assumptions." One can list several models that accurately recreate the power-law structure of citations, co-authorship, and research interest (Barabasi et al. 2002; Simkin / Roychaudhury 2007; Lafond 2015). Nonetheless, for the current contribution, we are particularly interested in random-walk models that forecast the growth of science (Shi et al. 2015; Jia et al. 2017; Iacopini et al. 2018). In other words, we take inspiration from earlier work that models scientists exploring a weighted network of people, ideas, and methods.

To construct our model, we also borrow from past research modeling online forums to study polarization and emotion contagion. Sobkowicz (2013: 3), for instance, offers a model consisting of agents and messages. In each round, the agents alternate between writing and reading each others' posts. And the author uses this simple ABM to create a "communication network from scratch using elements aimed at reproducing the flow of events in the real world." Therefore, we apply a similar methodology to a context where scientists exchange letters to express their findings and beliefs. Succinctly, we adjust their model to study historical letter correspondence.

Socio-Epistemic Networks

The basic setup for the model is a socio-epistemic network. On the social layer, scientists connect as senders and receivers of letters. And on the epistemic layer, we link ideas based on the co-occurrence of themes or the semantic similarity across letters. At last, agents also connect to topics/ideas based on their interest or the content of their previous letters. Altogether, we create a weighted multi-layered network whose edges represent the strength of association between scientists and ideas.

We create the socio-epistemic networks from the accumulated exchanged letters. Then, the ABM follows as an ego-reinforcement random walk across this network (Iacopini et al 2018; 2020). At each step, agents write new letters by exploring their local connections and choosing a receiver and a collection of topics. Hence, we use the weights of the edges to determine the chance that one agent picks a given receiver or theme. In turn, we reweight the edges at each step by the walker/sender. Thus, the socio-epistemic graph evolves by reinforcing its weights according to each new letter.

Model Summary

The model seeks to reproduce, grow, or generate artificial socio-epistemic networks representing the exchange of letters between agents. Thus, we propose a configuration model for the correspondence networks akin to a preferential attachment, small-world, or Erdos-Renyi model. But we ground our analysis on an agent-based simulation where, at each step, the agents must decide the content and the receivers of their letters. Along these lines, while our global structure follows a socio-epistemic network, it helps to describe the local agent decision as following a process similar to an extended Polya’s urn (Tria et al 2014; Casiraghi / Nanumyan 2021).

We start the ABM by creating agents and placing them into a geographical space. Then we give them a list of topics - each represented by a unique triple. And we assign them a social network or a list of people they have contacted in the past. Hence, every agent has a unique position, and they have two vectors keeping track of every letter sent and received. Besides, we give them an extra vector containing all their topics. Together, these vectors form a global socio-epistemic network, and make up the urns from which we pick the contents of the agent's future letters.

Agents have different activation rates. They have different propensities to initiate a conversation or send a letter. And when active, they first decide who to contact. Along these lines, we get a list of potential neighbors within a threshold radius. We consider social and spatial distance to make the list of neighbors, and the agent randomly selects and writes a letter to one of them. To pick the neighbor, we use a variation of Polya's urn. We create a distribution of ids based on their total letters received - their popularity - and past communication - both letters sent and received - with the focal agent. And we randomly pick one id from the distribution.

After we match senders with potential receivers, we compare their topic interests. We find the distance between their vectors of topics. And if it falls within a predetermined threshold, we continue with the simulation. Every letter has at least one topic. And we repeat the urn process to select it. Thus, we create a distribution of issues based on the intersection of interest between the receiver and sender, and we randomly pick one. After deciding on a receiver and topics, the agent can send their letters. And we need to update their global and agent variables.

After every letter, we update a global time-stamped ledger table that keeps track of every document sent. Each entry includes information on senders and receivers, their locations, the topic of the letter, and the time. So, we write the complete "transaction" information. In other words, we add a new line containing the sender's and receiver's ids, their respective locations, the topic, and the time.

Calibrating the Model

Our data includes letters from Central Europe around the 17th and 18th centuries. It contains available digital letters curated from several partner projects, such as LetterSampo, ePistolarium, Skillnet, and the correspSearch. Hence, our corpus corresponds to the aggregate data from all these projects, and we use it to train and test our Agent-Based Model. However, the minor details regarding each sample vary slightly from case to case, so we must adjust the model's initial conditions and parameters to fit our data input. For example, when focusing on topics or the semantic content of letters, we must rely primarily on the methods set by the LetterSampo project.

There are two top options for constructing the artificial correspondence network. First, like many earlier random walk models that predict the growth of science, we can apply the Republic of Letter as an input. We can collect every letter within a time window and use them to set up the model. And we execute the ABM to reconstruct the next 10 or 20 years of letter correspondence. In other words, we wish to use the out-of-sample prediction to test the ABM parameters and understand how well it reproduces real-world data. More importantly, we can use this approach to calibrate the parameters - e.g., the reinforcement weighting of the edges or how often agents write new letters.

Nonetheless, when using such methods, "there is a danger that the simulation would be trivial and reproduce the particular behavior without providing an insight into the generality of the mechanism involved" (Sobkowicz 2013: 2). Therefore, another complementary approach is to grow entirely new artificial letter networks – which take unique parameters and constraints.

For instance, from our available data, we can calculate the distribution of people in space, the number of letters sent and received by the agents, the most relevant topics, etc. Then, we can use this information to set up a small batch initial population following the patterns observed. Besides, we can add new cohorts of agents and ideas mimicking the exponential growth observed in many sources (Simkin / Roychowdhury 2007). At last, we can use a preferential attachment mechanism along with the ego-reinforcing random walk to evolve the correspondence network from scratch while still using our data sources to calibrate the ABM.

Model Outputs

The key idea behind the model is to grow artificial networks and use these as a benchmark, a null model to compare with the data and draw insights about historical correspondences and its archival. The model keeps track of every letter sent and received in this artificial environment. So, it outputs a "ledger" containing information about the letters - e.g., senders, receivers, addresses, and topics.

Furthermore, we can edit these to create a counterfactual database of surviving letters. For example, we can remove from the table all letters from minor agents - e.g., lesser-known people with fewer than 100 letters received. We can test what happens if we delete 10% or 25% of all letters, or we can remove all files from one place and time. We wish to measure the robustness of the network measures due to uncertainty and missing information - common issues when dealing with historical data.

Discussion

The model detailed above can contribute to the literature in different ways. First, we can use it to identify the "effect of data incompleteness on the relevant network measures" (Barabasi et al. 2002). The simulation aims to enhance our knowledge of the collection process behind the Republic of Letters. A common concern regarding historical letter data is their coverage - not all documents survive, and the ego networks usually center around a few notable individuals. Along these lines, when we remove letters from and to lesser-known agents from the ledger output, we aim to measure, at least partially, this coverage issue.

Past literature used methods from biology, like the unseen species model, to measure the size of missing documents or actors from historical resources (Kestemont et al. 2022; Wevers et al. 2022). Our contribution follows a similar direction, yet we do things slightly differently. Instead, we use an artificial society not to measure the amount of missing information but to assess how these lacking documents might impact our image of the past. We focus on historical letters and socio-epistemic networks to measure how robust our representation of past relationships is due to the missing data. We test how much the networks change in response to losing 10% or 30% of the available records.

Besides, calibrating our model against the Republic of Letters will substantiate our knowledge about the misrepresentation of the observed data. As we compare the empirical observations to a distribution of millions of alternative networks - about which we know the built-in parameters and details - we can learn about the process driving the creation, accumulation, and preservation of letters. We can, at least partially, measure the size of the sample bias and its origins. Take, for example, the underrepresentation of women in many historical sources. In our model, we can add parameters and artificially change the percentage of women in the sample, their likelihood of writing letters, and even the survivability rate of their letters. In doing so, we can create millions of alternative observations and use these to infer the data biases and their influence on our results.

The ABM can also serve as a configuration model, which we use to “discern observed features that can be expected at random from those beyond such expectations” (Casiraghi / Nanumyan 2021). In other words, we use the ABM to simulate counterfactual worlds. And these could lead to a better understanding of the factors that cause the network configurations we observe on the data.

By the same token, we can use these artificial networks to "normalize" or make the many existing social networks derived from existing correspondence data somewhat comparable. The Republic of Letters compiles many "isolated" case studies, and creating an artificial null model will serve as the foundation to generalize and compare their insights. Indeed, scholars often depend on null models like the Erdos-Renyi random graph to scale and evaluate the statistics derived from their multiple sources (Anderson et al. 1999; Faust / Skvoretz, 2002; van Wijk et al. 2010).

At last, currently, we focus the model on understanding the biases and the robustness of existing practices graphing historical networks. We want to learn more about the uncertainty around data coverage. Hence, we are not yet ready to examine the dynamics of these socio-epistemic networks. These are valuable research questions and chief concerns of the larger project of which this proposal is part - the ModelSEN. We do wish to investigate these matters in the future. Yet, to do so, we first need a successful model that can reproduce communication in the Republic of Letters. Once we have developed the ABM, we can start thinking about the diffusion of ideas through historical correspondences (Rossini 2022) or the influences of disenfranchisement on the epistemic network. Much like Sobkowicz (2013) used the communication process detailed here to grow their network and later simulate the dynamics of opinions across the artificial population, one could employ our model of communication during the Republic of Letters as the background for examining how ideas - e.g., Decarte's ideas - diffuse from one writer to another as they communicate through letters. Alternatively, like the gender example from before, we can tune the model parameters. We can artificially change the population size and structure to see how these influence knowledge dynamics. In other words, if we can effectively simulate historical correspondences, we could, in the future, simulate how our agents learn, adapt, and write about new ideas and topics.

Bibliography

Anderson, Brigham S. / Butts, Carter / Carley, Kathleen (1999): “The interaction of size and density with graph-level indices”, in: *Social Networks* 21, 3: 239-267.

Barabasi, Albert-Laszlo / Jeong, Hawoong / Neda, Zoltan / Ravasz, Erzsebet / Schubert, Andreas / Vicsek, Tamas (2002): “Evolution of the social network of scientific collaborations”, in: *Physica A: Statistical Mechanics and Its Applications* 311, 3-4: 590-614.

Casiraghi, Giona / Nanumyan, Vahan (2021): "Configuration models as an urn problem", in: *Scientific Reports* 11, 1: 13416.

Edelstein, Dan / Findlen, Paula / Caserani, Giovanna / Winterer, Caroline / Coleman, Nicole (2017): “Historical research in a digital age: Reflections from the mapping the Republic of Letters project”, in: *The American Historical Review 122* , 2: 400-424.

Faust, Katherine / Skvoretz, John (2002): “Comparing networks across space and time, size and species”, in: *Sociological Methodology* 32, 1: 267-299.

Gilbert, Nigel (1997): “A simulation of the structure of academic science”, in: Sociological Research Online 2, 2: 91-105.

Hotson, Howard / Walling, Thomas (2019): *Reassembling the Republic of Letters in the Digital Age* . G ö ttingen: G ö ttingen University Press.

Iacopini, Iacopo / Milojević, Staša Milojević / Latora, Vito (2018): “Network dynamics of innovation processes”, in: *Physical Review Letters* 120, 4: 048301.

Iacopini, Iacopo / Di Bona, Gabriele / Ubaldi, Enrico / Loreto, Vittorio / Latora, Vito (2020): “Interacting discovery processes on complex networks”, in: *Physical Review Letters* 124, 24: 248301.

Jia, Tao / Wang, Dashun / Szymanski, Boleslaw K. (2017): “Quantifying patterns of research-interest evolution”, in: *Nature Human Behavior* 1, 4: 0078.

Kestemont, Mike / Karsdorp, Folgert / de Bruijn, Elisabeth / Driscoll, Matthew / Kapitan, Katarzyna A. / Ó Macháin, Pádraig / Sawyer, Daniel / Sleiderink, Remco / Chao, Anne (2022): "Forgotten books: The application of unseen species models to the survival of culture", in: *Science* 375, 6582: 765-769.

Lafond, François (2015): “Self-organization of knowledge economies”, in: *Journal of Economic Dynamics and Control* 52: 150-165.

Rossini, Paolo (2022): “The networked origins of cartesian philosophy and science”, in: *The Journal of the International Society for the History of Philosophy and Science* 12, 1 *:* 97-120.

Shi, Feng / Foster, Jacob G. / Evans, James A. (2015): “Weaving the fabric of science: Dynamic network models of science’s unfolding structure”, in: *Social Networks* 43: 73-85.

Simkin, Mikhail V. / Roychowdhury, Vwani P. (2007): “A mathematical theory of citing”, in: *Journal of the American Society for Information Science and Technology* 58, 11: 1661-1673.

Sobkowicz, Pawel (2013): “Quantitative agent based model of user behavior in an internet discussion forum”, in: *PLoS ONE* 8, 12: e80524.

Tria, Francesca / Loreto, Vittorio / Servedio, Vito D. P. / Strogatz, Steven H. (2014). “The dynamics of correlated novelties”, in: Scientific Reports 4, 1: 1-8.

Ureña-Carrion, Javier / Leskinen, Petri / Tuominen, Jouni / van den Heuvel, Charles / Hyvönen, Eero / Kivel ä Mikko (2022): “Communication now and then: Analyzing the Republic of Letters as a communication network”, in: *Applied Network Science* 7, 26.

van den Heuvel, Charles (2015): “Mapping knowledge exchange in early modern Europe: Intellectual and technological geographies and network representations”, in: International Journal of Humanities and Arts Computing 9, 1: 95-114.

van Wijk, Bernadette C. M. / Stam, Cornelis J. / Daffertshofer, Andreas (2010): “Comparing brain networks of different size and connectivity density using graph theory”, in: *PLoS One 5*, 10: e13701.

Wever, Melvin / Karsdorp, Folgert / van Lottum, Jelle (2022): “What shall we do with the unseen sailor? Estimating the size of the Dutch East India Company using an unseen species model.” in: *CHR 2022: Computational Humanities Research Conference* , Antwerp, Belgium, December 2022: 189 197.