The diffusion of scientific terms – tracing individuals' influence in the history of science for English

Yuri Bizzoni and **Stefania Degaetano-Ortlieb** and **Katrin Menzel** and **Elke Teich**
Department of Language Science and Technology, Saarland University
yuri.bizzoni, s.degaetano, k.menzel, e.teich@mx.uni-saarland.de

Abstract

Tracing the influence of individuals or groups in social networks is an increasingly popular task in sociolinguistic studies. While methods to determine someone's influence in short-term contexts (e.g., social media, on-line political debates) are widespread, influence in long-term contexts is less investigated and may be harder to capture. We study the diffusion of scientific terms in an English diachronic scientific corpus, applying Hawkes Processes to capture the role of individual scientists as "influencers" or "influencees" in the diffusion of new concepts. Our findings on two major scientific discoveries in chemistry and astronomy of the 18th century reveal that modelling both the introduction and diffusion of scientific terms in a historical corpus as Hawkes Processes allows detecting patterns of influence between authors on a long-term scale.

1 Introduction

The detection and definition of influence in social contexts is becoming increasingly popular in computational sociolinguistics and related disciplines (Eisenstein et al., 2014; Nguyen et al., 2016; Cork et al., 2020), mainly due to the availability of data from large-scale social media that allow us to study an unprecedented amount of exchanges, but recently also in diachronic settings as shown by Soni et al. (2021) on lexical semantic change. As the diffusion of new concepts and ideas is inherent to scientific writing, one can assume many cases of terminological influence between authors. Also, since scientific communities act as social and collaborative networks (Newman, 2001), the diffusion of new concepts in the history of science might behave similarly to that of recently formed communities such as social media networks. Scientific findings that either introduce new concepts or redefine the understanding of known phenomena are bound to either encourage the acquisition of new words and expressions or to promote existing terminology that gets endowed with novel importance.

We present a pilot application of Hawkes Processes, usually used to trace short-term influences between cascades of events, modeling long-term influences between authors of the Philosophical Transactions and Proceedings of the Royal Society of London, two of the longest-running scientific journals in English. From today's perspective, the 18th century Royal Society network represented a rather narrow circle of professional and private relations (Røstvik and Fyfe, 2018). In the publications of this "scientific club", we assume a reflection of social and thematic networks. Central members may have had a particular influence by crossing disciplinary boundaries that were not yet well established. Relying on previous findings on innovative uses of words in the Late Modern Period of the RSC (Degaetano-Ortlieb and Teich, 2019), we select in a data-driven fashion using Kullback-Leibler Divergence sets of words distinctively used around that period and involved in terminology formation. These sets of words are used to select texts to be fed into the modelling of Hawkes Processes. We present the results of influences between individuals for two specific "revolutionary" periods in the history of science: (1) the discovery of oxygen, leading to the Chemical Revolution in the mid-18th century, and (2) the rapid expansion of knowledge on the composition of structures in the universe (the planetary nebula) through more powerful telescopes around the end of the 18th century.

2 Related Work

Large diachronic scientific corpora are rare and valuable, since they allow us to study the development of a rather big but specialized community throughout historical time spans. One of the largest curated corpora of this kind to date is the Royal Society Corpus (RSC; see Kermes et al. (2016); Fischer et al. (2020) for corpus creation and Menzel

et al. (2021) for metadata curation). The material of the Royal Society has been used in various diachronic studies to explore the development of English scientific writing and aspects such as the evolution of peer review and community-based editorial processes of scholarly journals (e.g., Halliday (1988); Degaetano-Ortlieb and Teich (2019); Fyfe et al. (2020); Bizzoni et al. (2020)).

Considering work on tracing influences, Hawkes Processes (HP) (Hawkes, 1971), originally designed to model waves of earthquakes' aftershocks (Hawkes, 1973), have gained popularity in sociolinguistics. They have been recently used as powerful means to study influence relationships between authors on a text-based level (He et al., 2015) as well as words' spreading and turn taking in conversations (Li et al., 2014; Guo et al., 2015; Daw et al., 2020) and social media (Goel et al., 2016; Zhang et al., 2018; Soni et al., 2019; Dutta et al., 2020).

The main advantage of using HPs has been found in their ability to model a wide range of different relationships between events through a transfer function (Bonnet et al., 2021). Also, their flexibility in accommodating auxiliary features (Chiang et al., 2021) has made them increasingly popular in fields like criminology (Mohler et al., 2018), economy (Rambaldi et al., 2015; Hawkes, 2018), and epidemiology (Escobar, 2020). Their flexibility has also made HPs particularly apt to finance and market modelling (Bacry et al., 2015; Hawkes, 2018). Naturally, with the growth of social media a number of studies have found HPs suitable to model the spread of information through communication networks (Lukasik et al., 2016; Rizoiu et al., 2017; Li and Bhowmick, 2020). All in all, HPs are a relatively simple and robust method to detect influence from incomplete and noisy data (Da Fonseca and Zaatour, 2014; Yang et al., 2018). We adapt HPs to model long-term influences between scientific authors focusing on the use of newly established terminology.

3 Data and Methods

3.1 The Royal Society Corpus

The Royal Society Corpus (RSC; Kermes et al. (2016); Fischer et al. (2020); Menzel et al. (2021)) is composed of the full set of publications from the period between 1665 and 1996 of the Royal Society of London. As multidisciplinary journals covering the natural sciences, the Philosophical Transactions and Proceedings of the Royal Society originally used to cover all branches of science of the time and became increasingly specialized due to increased competition from commercial journals and publications from specialized scientific societies. Few journals are so deeply embedded in a community of scholars as these (Røstvik and Fyfe, 2018).[1] We use the open access RSC 4.0 version[2] (1665-1869, 9.779 texts, 31.952.725 tokens) as it covers the period where we previously detected major scientific discoveries reported in the RSC such as the Chemical Revolution (Degaetano-Ortlieb and Teich, 2019). The RSC encodes metadata (such as text type (article, abstract), author, title, date of publication, and time periods (decades, fifty years)) and linguistic annotation at the levels of tokens (with normalized and original forms), lemmas, and parts of speech using TreeTagger (Schmid, 1995), achieving 95.1% accuracy on normalized word forms (normalization is based on VARD; see Baron and Rayson (2008)). As the material has been OCRed, pre-processing steps have been adopted to mitigate OCR errors as good as possible (see Kermes et al. (2016) and recently Menzel et al. (2021) for a more detailed description).

3.2 Modeling Event Cascades by Hawkes Processes

A Hawkes Process (Hawkes, 1971) (HP) is a self-exciting point process, where the probability of seeing new events increases based on previous events. This is a way of modeling event cascades, where some events seem to cause or "excite" a number of successive events (consider e.g. an earthquake and its after-shocks). In simpler terms, they are sequences of events happening with some form of domino effect. For example, the first shock of an earthquake happens randomly, while the succession of after-shocks comes only as a consequence of the first one. Since the first shock initiates a cascade of events, the events that immediately follow it will not show an arbitrary distribution. In linguistic exchanges, such as conversations or letters, HPs can be applied to model influence among interactants and be also applied to capture the diffusion of new

---

[1]The Royal Society of the 17th-19th century was a rather closed circle. Its journals became fundamentally linked to the Society's scientific meetings. Manuscripts could only be submitted by fellows of the RS, a professional and social network dedicated to discussing science and investigating facts via scientific experiments. Fellows could also "communicate" manuscripts for nonmembers.

[2]RSC 4.0 at https://hdl.handle.net/21.11119/0000-0001-7E8B-6 and further releases.

words or linguistic uses, e.g. in a debate or social network (Linderman and Adams, 2014; Soni et al., 2019).

We consider words as events over the time line. If author X introduces a new word, and Y and Z use it in a pattern that resembles that of an event cascade rather than a random occurrence, we may assume an influence from X to Y and Z.[3] Since its aim is to understand who influences whom in a network, HP has an event intensity function (cf. Rizoiu et al. (2017)):

$$\lambda_j(t) = \lambda_{j,0} + \sum_{t' < t} \alpha_{s_e \to j} \kappa(t - t') \qquad (1)$$

where α is a scalar excitement parameter, s_e indicates the source of event e (here: a word), $\alpha_{s_e \to j}$ quantifies how much the source event s_e tends to excite target j, and κ is a kernel decay function monotonically decreasing through time, constrained to integrate to 1 over positive arguments – the further away in time, the weaker the influence is expected to be. The model's overall goal is to measure the intensity of the influence of s_e over j for a specific time interval Δt and is modeled as a sum over B simple basis models (here B is equal to the number of authors considered):

$$\kappa_{s_e \to j}(\Delta t) = \sum_{b=1}^{B} g_b^{(s_e \to j)} \phi_b(\Delta t). \qquad (2)$$

where $\phi_b(\Delta t)$ is the basis model (an impulse function that sums to 1) and g_b are the dyad-specific weights over the basis models.

As input for the Hawkes Process model[4], we use sentences (as plain text) from the RSC corpus which contain sets of words attributed to major discoveries. The selection procedure of these sets of words is described in the following section. To obtain an overall impression of influencer and influencees, we use a heatmap visualization based on Equation 2 showing the intensity of influence across groups.

3.3 Selection Procedure

For selection of words as possible events which might trigger cascades from an influencing author to influencees, rather than intuitively choosing words, we use Kullback-Leibler Divergence (KLD) (Kullback and Leibler, 1951) as a more data-driven approach, modeling at a given time period differences between past and future language use following Degaetano-Ortlieb and Teich (2018). Formally, we compare the probability of a word w in a period of 20-years following a given year, $p(w|F)$, to the word's probability in a period of 20-years preceding a given year, $p(w|P)$. Given these probabilities, we model divergence between future language use F given past language use P:

$$D(F||P) = \sum_i p(w_i|F) log_2 \frac{p(w_i|F)}{p(w_i|P)} \qquad (3)$$

The words with the highest contribution to the overall KLD at a particular point in time denote words used distinctively in the future compared to the past, i.e. new words or existing words used in novel ways. From this ranking, we observe words related to two major discoveries reported in the RSC corpus. First, we consider the *discovery of oxygen* in 1774 by Joseph Priestley, which substituted the phlogiston theory of combustion. This transition was characterized by the coinage and spreading of words related to the new element. Second, we consider the *discovery of new nebulae* (mainly the planetary nebulae) at the end of the 18th century by William and Caroline Herschel. This expansion of scientific knowledge was characterized by the increased success of existing terminology (like the word *nebula* itself). We select the most distinctive words for each discovery from the KLD rankings.[5] We use these word sets to extract sentences from the RSC corpus, the respective authors and the publication year. This data serves as input for the Hawkes Process model.

For both discoveries (oxygen and nebula), in this pilot study, we limit our analysis to single-author papers of authors who published more than seven papers containing at least one of the words of our selection.[6]

---

[3] Note that the concept of influence is to be held as likely, and not absolutely ascertained. We are de facto observing correlation, not causation. For example, it may be that all parties are being influenced by the same hidden source with a delay, i.e. early adopters would not be influencing late adopters. Saying that X influences Y means that their behavior is consistent with that of an event cascade initiated by X and continued by Y.

[4] https://github.com/degaetano-ortlieb/influencer_RSC _HawkesProcess.

[5] oxygen set: *oxyg(en|enous|ene), phlogiston, dephlogisticated, acid, water, gas, hydrogen(e), substance, solution, oxide, compound, muriatic*; nebula set: *nebul(a|ae|e|osity|ous|ar|osa|as), Saturn, luminous, bright(ness), illuminate, position, cluster, object, lustre, spectrum, eye, star, ray, light, catalogue, formation, rotation*

[6] Note that in the RSC, there are some non-unified au-

4 Influencers and Influencees in the RSC

4.1 The Discovery of Oxygen

Figure 1 shows an influencer heatmap of the most prolific authors during the diffusion of the theory of oxygen. The y-axis denotes the influencers and
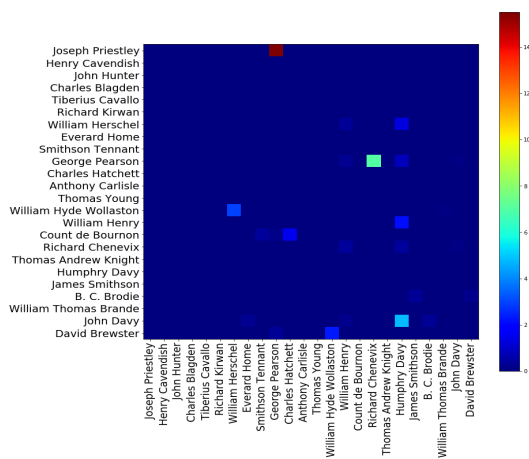

Figure 1: Influencer heatmap (oxygen)

the x-axis the influencees. Stronger influences shift towards the red. The single strongest influence is by Joseph Priestley on George Pearson: after Priestley's first use of words such as *oxygenous* (see Figure 2), Pearson follows with a series of mentions of morphologically related words which our model detects as a highly probable cascade reaction to Priestley.[7] In fact, Pearson was the first English chemist to adopt Priestley's oxygen theory (Coley, 2003), and while Priestley displays the strongest influence of the group, Pearson's work has the most widespread effect, showing direct influence on eleven other scientists we consider, including Chenevix, Davy, and Henry (see Appendix Table 2). Additionally, some authors are influenced by several others, e.g., Humphry Davy (the first to describe the properties of nitrous oxide, a gas still extensively used in anesthesia) by Herschel, Pearson, Henry, Chenevix, and John Davy.

Overall, it seems that oxygen's diffusion happens through three main roles which can be partly re-

lated to the Diffusion of Innovation theory (Rogers, 2010): the first user, Priestley, with a strong but focalized influence acts as an *innovator*; Pearson, an influencee-influencer acts as an *early adopter*, reacting strongly to Priestley and in turn influencing a large number of peers; and strong influencees like Davy, reacting to an already widespread use from several authors with a large number of mentions possibly acting as *early majority* adopters.[8]
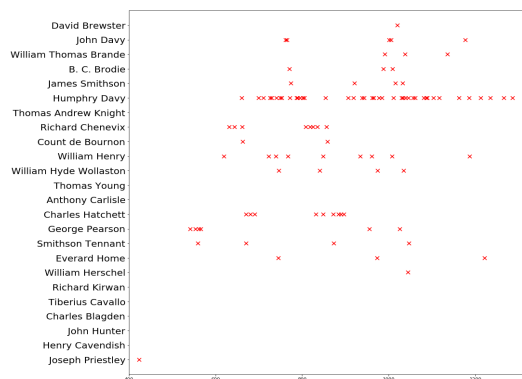

Figure 2: Chronologically ordered mentions of the word *oxygen*

4.2 The Discovery of the Planetary Nebula

Figure 3 shows the most prolific authors using the set of words we selected for the astronomical field in the period between 1761 and 1828. William Herschel is visibly influenced by many other authors. His "big influencee" role seems to reflect his research endeavor of collecting/discussing facts in the community (e.g., Francis Wollaston's catalogue or the relationship to his friend Pigott (Hoskin, 1979)). Note that his strongest influencer was his sister Caroline Herschel (see red square in Appendix Figure 4).[9] The discovery of new instances of pre-existing "objects" like nebulae ultimately lead him to discover the planetary nebula as a different phenomenon from star clusters. William Herschel did not influence this sub-field with new terminology, but did "resonate" to concepts already in the field and mentioned by others.

Beyond the Herschels, this group of astronomers and physicists reveals other valid influences. For example, Henry Kater, famous for his studies on

---

thor names, i.e. while most authors are covered by dominant spellings, some have several spelling variants, which are however rare. For practical reasons, in this pilot work, we consider only the most frequent spelling variants of the respective author names. Currently, to overcome this issue, unique Fellow IDs for authors that were fellows of the RS are being added to the corpus metadata, allowing us in future work to further increase coverage of spelling variants.

[7]Priestley himself did not use the word oxygen(e) yet, which was coined in France by Lavoisier, who learned the process of making oxygen from Priestley.

---

[8]The role of Davy as a prolific inventor and President of the RS from 1820 could also be a factor in his behaviour.

[9]Caroline was not a prolific author and does not appear in Figure 3. However, based on her well-known contribution to her brother's work, we added her texts for inspection revealing her strong influence on her brother.

telescopes, has a relevant influence over the pioneer optician and inventor David Brewster (see red square in Figure 3 and Table 3) or Edward Sabine, actively calling for building magnetic observatories, influences Samuel Hunter Christie pursing the same interest (MacDonald, 2015, 416) (orange square in Figure 3 and Table 3).
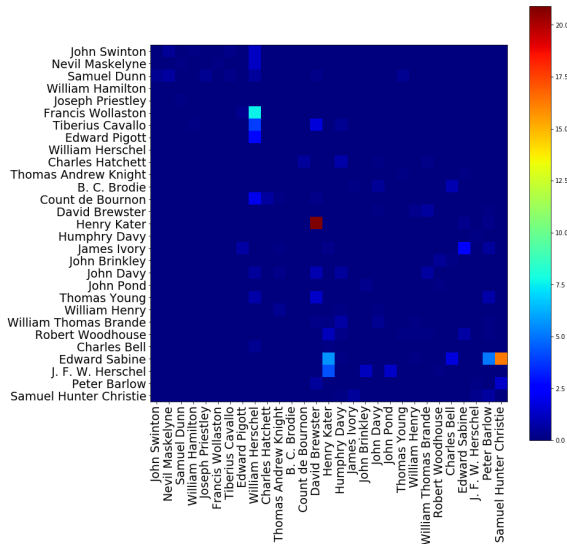

Figure 3: Influencer heatmap (nebula)

Overall, rather than the early adopter role acting as strong influencer as seen for the diffusion of oxygen, here there are writers who start using existing terminology with increased insistence as the influencee William Herschel, wide-spreading influencer as Caroline Herschel and others (see Table 4), and strong individual influencers as Kater on Brewster.

5 Discussion and Future work

This pilot study aimed to trace influences on the diffusion of terms in the history of science. We model long-term influences with Hawkes Processes, usually applied to short-time studies, on the diachronic Royal Society Corpus. We focus on two cases in the Late Modern Period: (a) the diffusion of oxygen terminology, when the competing concept of phlogiston was losing momentum, and (b) the diffusion of specific astronomical terms centered around the discovery of new nebulae. For both, the model detects realistic influences between authors which correspond to qualitative observations of the scientific fields involved. The relationships between influencers and influencees can be quite diverse and contribute differently to the diffusion of innovative knowledge. For example, for the oxygen theory, while Priestley is the innovator, Pearson, the early

adopter, is the one influencing a large number of peers. For the planetary nebulae discovery, instead, Caroline Herschel acts as a wide-spreading influencer among other influencers, while her brother William is a strong adopter with no major influencer role. Being supported by empirical observations, these results also validate HPs as a possible tool to study long-term influence between authors.

In future works, a deeper qualitative analysis will allow us to detect more reliably possible benefits and drawbacks of the approach. For example, informing our data-driven approach with knowledge of historians of science, as we have simplistically shown for Caroline Herschel, would be beneficial, and investigating more deeply connections detected by our model will show whether we are able to find latent patterns of influence or detect false positives. To further evaluate the appropriateness of the approach, we might attempt to reproduce the phenomena we model in synthetic data that mimic types of lexical diffusion over longer periods. Thus, we would have an even clearer idea of what types of scientific influence HPs can capture best.

While we here consider the lexical level, we are especially interested in tracing the diffusion of syntactic and/or stylistic patterns shaping the scientific register. The language of science has been shown to increasingly use conventionalized ways of expression to level out the high degree of informativity rendered (Biber and Gray, 2016; Degaetano-Ortlieb and Teich, 2019; Teich et al., 2021). However, influencers originating from minority groups, such as Caroline Herschel, might have had to oblige not only to conventions related to the language of science but also to socio-cultural conventions possibly reflected in their use of language. According to Winterburn (2015, 3,7,8), e.g., Caroline's language is modest, self-effacing and polite highlighting her awareness of the proper codes of conduct for women, but also succinct and rigorous on the accounts of her observations, conclusions and defence of their credibility adhering to the expectations of the scientific community. Also, her use of language does not claim priority as her bother's, but focuses on sharing information. Thus, detecting influencer belonging to particular socio-cultural groups and analyzing their use of language across linguistic levels might help us to better understand the complexity of the diffusion of knowledge and the different possible stages and relationships of influencers and influencees considering social roles.

References

Emmanuel Bacry, Iacopo Mastromatteo, and Jean-François Muzy. 2015. Hawkes processes in finance. *Market Microstructure and Liquidity*, 1(01):1550005.

Alistair Baron and Paul Rayson. 2008. VARD 2: A tool for dealing with spelling variation in historical corpora. In *Proceedings of the Postgraduate Conference in Corpus Linguistics*, Birmingham, UK.

Douglas Biber and Bethany Gray. 2016. *Grammatical complexity in academic English: Linguistic change in writing*. Studies in English Language. Cambridge University Press, Cambridge, UK.

Yuri Bizzoni, Stefania Degaetano-Ortlieb, Peter Fankhauser, and Elke Teich. 2020. Linguistic variation and change in 250 years of English scientific writing: A data-driven approach. *Frontiers in Artificial Intelligence*, 3:73.

Anna Bonnet, Miguel Herrera, and Maxime Sangnier. 2021. Maximum likelihood estimation for Hawkes processes with self-excitation or inhibition. *arXiv preprint arXiv:2103.05299*.

Wen-Hao Chiang, Xueying Liu, and George Mohler. 2021. Hawkes process modeling of covid-19 with mobility leading indicators and spatial covariates. *International Journal of Forecasting*, online print.

Noel G. Coley. 2003. George Pearson MD, FRS (1751-1828): The greatest chemist in England? *Notes and Records of the Royal Society of London*, 57(2):161–175.

Alicia Cork, Richard Everson, Mark Levine, and Miriam Koschate. 2020. Using computational techniques to study social influence online. *Group Processes & Intergroup Relations*, 23(6):808–826.

José Da Fonseca and Riadh Zaatour. 2014. Hawkes process: Fast calibration, application to trade clustering, and diffusive limit. *Journal of Futures Markets*, 34(6):548–579.

Andrew Daw, Antonio Castellanos, Galit B Yom-Tov, Jamol Pender, and Leor Gruendlinger. 2020. The co-production of service: Modeling service times in contact centers using Hawkes processes. *arXiv preprint arXiv:2004.07861*.

Stefania Degaetano-Ortlieb and Elke Teich. 2018. Using relative entropy for detection and analysis of periods of diachronic linguistic change. In *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 22–33, Santa Fe, New Mexico. Association for Computational Linguistics.

Stefania Degaetano-Ortlieb and Elke Teich. 2019. Toward an optimal code for communication: The case of scientific English. *Corpus Linguistics and Linguistic Theory*, 0(0):1–33. Online print.

Hridoy Sankar Dutta, Vishal Raj Dutta, Aditya Adhikary, and Tanmoy Chakraborty. 2020. Hawkes-Eye: Detecting fake retweeters using Hawkes process and topic modeling. *IEEE Transactions on Information Forensics and Security*, 15:2667–2678.

Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. 2014. Diffusion of lexical change in social media. *PLOS ONE*, 9(11):1–13.

Juan V. Escobar. 2020. A Hawkes process model for the propagation of COVID-19: Simple analytical results. *Europhysics Letters (EPL)*, 131(6):68005.

Stefan Fischer, Jörg Knappen, Katrin Menzel, and Elke Teich. 2020. The Royal Society Corpus 6.0. Providing 300+ years of scientific writing for humanistic study. In *Proceedings of the the 12th Language Resources and Evaluation Conference (LREC)*, Marseille, France.

Aileen Fyfe, Flaminio Squazzoni, Didier Torny, and Pierpaolo Dondio. 2020. Managing the growth of peer review at the Royal Society journals, 1865-1965. *Science, Technology, and Human Values*, 45(3):405–429.

Rahul Goel, Sandeep Soni, Naman Goyal, John Paparrizos, Hanna Wallach, Fernando Diaz, and Jacob Eisenstein. 2016. The social dynamics of language change in online networks. In *International Conference on Social Informatics*, pages 41–57. Springer.

Fangjian Guo, Charles Blundell, Hanna Wallach, and Katherine Heller. 2015. The Bayesian Echo Chamber: Modeling social influence via linguistic accommodation. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38 of *Proceedings of Machine Learning Research*, pages 315–323, San Diego, California, USA. PMLR.

M.A.K. Halliday. 1988. On the language of physical science. In Mohsen Ghadessy, editor, *Registers of Written English: Situational Factors and Linguistic Features*, pages 162–177. Pinter, London.

Alan G Hawkes. 1971. Point spectra of some mutually exciting point processes. *Journal of the Royal Statistical Society: Series B (Methodological)*, 33(3):438–443.

Alan G. Hawkes. 1973. Cluster models for earthquakes-regional comparisons. *Bull. Int. Stat. Inst.*, 45(3):454–461.

Alan G Hawkes. 2018. Hawkes processes and their applications to finance: a review. *Quantitative Finance*, 18(2):193–198.

Xinran He, Theodoros Rekatsinas, James Foulds, Lise Getoor, and Yan Liu. 2015. Hawkestopic: A joint model for network inference and topic modeling from text-based cascades. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 871–880, Lille, France. PMLR.

Michael Hoskin. 1979. Goodricke, Pigott and the quest for variable stars. *Journal for the History of Astronomy*, 10:23–24.

Hannah Kermes, Stefania Degaetano-Ortlieb, Ashraf Khamis, Jörg Knappen, and Elke Teich. 2016. The Royal Society Corpus: From uncharted data to corpus. In *Proceedings of the 10th LREC*, Portorož, Slovenia.

Solomon Kullback and Richard A. Leibler. 1951. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

Hui Li and Sourav S Bhowmick. 2020. Brunch: branching structure inference of hybrid multivariate Hawkes processes with application to social media. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 553–566. Springer.

Liangda Li, Hongbo Deng, Anlei Dong, Yi Chang, and Hongyuan Zha. 2014. Identifying and labeling search tasks via query-based Hawkes processes. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, page 731–740, New York, NY, USA. Association for Computing Machinery.

Scott Linderman and Ryan Adams. 2014. Discovering latent network structure in point process data. In *Proceedings of the 31st International Conference on Machine Learning*, pages 1413–1421. PLMR.

Michal Lukasik, PK Srijith, Duy Vu, Kalina Bontcheva, Arkaitz Zubiaga, and Trevor Cohn. 2016. Hawkes processes for continuous time sequence classification: an application to rumour stance classification in twitter. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 393–398.

Lee T. MacDonald. 2015. Making kew observatory: the Royal Society, the British Association and the politics of early Victorian science. *The British Journal for the History of Science*, 48(3):409–433.

Katrin Menzel, Jörg Knappen, and Elke Teich. 2021. Generating linguistically relevant metadata for the Royal Society Corpus. *Challenges in Combining Structured and Unstructured Data in Corpus Development - Special Issue of Research in Corpus Linguistics, Asociación Española de Lingüística de Corpus (AELINCO)*, 9(1):1–18.

George Mohler, Jeremy Carter, and Rajeev Raje. 2018. Improving social harm indices with a modulated Hawkes process. *International Journal of Forecasting*, 34(3):431–439.

Mark E.J. Newman. 2001. Scientific collaboration networks. I. Network construction and fundamental results. *Physical Review E*, 64(016131).

Dong Nguyen, A. Seza Doğruöz, Carolyn P. Rosé, and Franciska De Jong. 2016. Computational sociolinguistics: A survey. *Computational Linguistics*, 42(3):537–593.

Marcello Rambaldi, Paris Pennesi, and Fabrizio Lillo. 2015. Modeling foreign exchange market activity around macroeconomic news: Hawkes-process approach. *Physical Review E*, 91(1):012819.

Marian-Andrei Rizoiu, Young Lee, Swapnil Mishra, and Lexing Xie. 2017. *Hawkes processes for events in social media*, page 191–218. Association for Computing Machinery and Morgan & Claypool.

Everette M. Rogers. 2010. *Diffusion of Innovation*. The Free Press, New York.

Camilla Mørk Røstvik and Aileen Fyfe. 2018. Ladies, gentlemen, and scientific publication at the Royal Society, 1945–1990. *Open Library of Humanities*, 4(1):1–40.

Sandeep Soni, Lauren F. Klein, and Jacob Eisenstein. 2021. Abolitionist networks: Modeling language change in nineteenth-century activist newspapers. *Journal of Cultural Analytics*, 6(1).

Sandeep Soni, Shawn Ling Ramirez, and Jacob Eisenstein. 2019. Detecting social influence in event cascades by comparing discriminative rankers. In *The 2019 ACM SIGKDD Workshop on Causal Discovery*, pages 78–99. PMLR.

Elke Teich, Peter Fankhauser, Stefania Degaetano-Ortlieb, and Yuri Bizzoni. 2021. Less is more/more diverse: On the communicative utility of linguistic conventionalization. *Frontiers in Communication*, 5:142.

Emily Winterburn. 2015. Learned modesty and the first lady's comet: A commentary on Caroline Herschel (1787) 'An account of a new comet'. *Philosophical Transactions of the Royal Society A*, 373(2039):20140210.

Steve Y Yang, Anqi Liu, Jing Chen, and Alan Hawkes. 2018. Applications of a multivariate Hawkes process to joint modeling of sentiment and market return events. *Quantitative Finance*, 18(2):295–310.

Wei Zhang, Fan Bu, Derek Owens-Oas, Katherine Heller, and Xiaojin Zhu. 2018. Learning root source with marked multivariate Hawkes processes. *arXiv preprint arXiv:1809.03648*.

A Appendix

Table 1: Strongest individual influences related to oxygen.

Influence Score	Authors
15.5	Joseph Priestley on George Pearson
6.98	George Pearson on Richard Chenevix
4.80	John Davy on Humphry Davy
3.51	William Hyde Wollaston on Humphry Davy
2.31	David Brewster on William Hyde Wollaston
2.15	William Henry on Humphry Davy

Table 2: Most wide-reaching influencers related to oxygen.

No. of Influenced Authors	Author
11	George Pearson
8	Richard Chenevix
6	William Herschel
4	John Davy
3	William Hyde Wollaston
3	Count de Bournon

Table 3: Strongest individual influences in the astronomical field.

Influence Score	Authors
169.42	Caroline Herschel on William Herschel
22.54	John Hunter on William Herschel
20.89	Henry Kater on David Brewster
16.28	Edward Sabine on Samuel Hunter Christie
7.14	John Ellis on Nevil Maskelyne
5.88	John Brinkley on Peter Barlow

Table 4: Most wide-reaching influencers in the astronomical field.

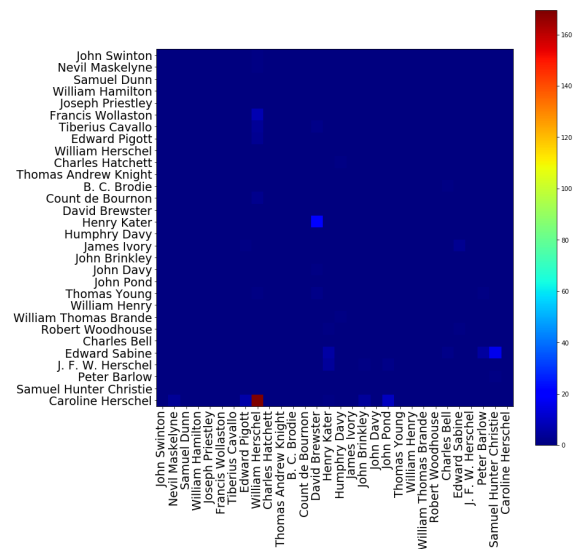| No. of Influenced Authors	Author
19	Caroline Herschel
17	Samuel Dunn
14	Robert Woodhouse
12	John Swinton
12	Count de Bournon
11	Thomas Young

Figure 4: Influencer heatmap (nebula) plus Caroline Herschel.