Embed, Detect and Describe: A Framework for Examining Events in Complex Sociocultural and Historical Data

Melvin Wevers[1*], Jan Kostkan[2], Kristoffer Nielbo[2]

[1] Dept. of History, University of Amsterdam, Netherlands
[2] Center for Humanities Computing, Aarhus University, Denmark
[*] Corresponding author, kln@cas.au.dk

A central building block for historical research is historical events, that is, dynamic objects displaced in time. Despite their importance, we see a disconnect between theoretical work and empirical studies of events [1]. This is exemplified by what we will refer to as the *Euclidean Error* in historical reconstructions. While historians generally agree that historical events are complex and non-linear in theory, empirical research is ripe with approaches that, due to data sparsity or inadequate formalization, describe history as consisting of singular dates, 'event change points' that are connected by uneventful lines, static 'event states' with low sensitivity to temporal variation, and, consequently, an overly reductive reconstruction of historical events. To counter this approach, we propose an alternative formal framework – *Embed, Detect and Describe* – an information-theoretical approach to (historical) event detection and description in noisy and complex sociocultural data. The framework is based on a fundamental theorem of chaos theory, the embedding theorem [2, 3, 4], which allows us to approximate the dynamics of a large-scale social system. Rather than measuring cultural expressions through, for instance, word counts over time, we approach society as a complex system with a multitude of states, which switch between attractors, i.e., a value or set of values toward which variables in a dynamical system tend to evolve. Some of these attractors may be associated with dynamics of cultural information and captured in low-dimensional indicator variables [5, 6]. In our case studies, see Fig. 1, these indicator variables are expressed through the amount of surprise encoded in the textual content of news media. By this, we mean how much of the information at one point in time can be expected given an earlier time point. If the data is almost the same, there is a low surprise; if it is radically different, the surprise increases. The framework is fundamentally data agnostic and will apply to any dense and low-rank embedding of the data objects, e.g., text, sound, or image, with some minor modifications. Importantly, our approach to events is psychological, i.e., we study how humans organize and understand events rather than attempt to formalize an event ontology [7]. The talk describes two techniques for detecting and describing changes between event states. Although these techniques rely on information theory and Bayesian inference, they are members of two generic sets of formal techniques for identifying and characterizing differences in the state of a process at different times. The specific choice of models and algorithms is secondary to the main argument, namely that digital historical research has to pay more attention to complexities involved with change and events.

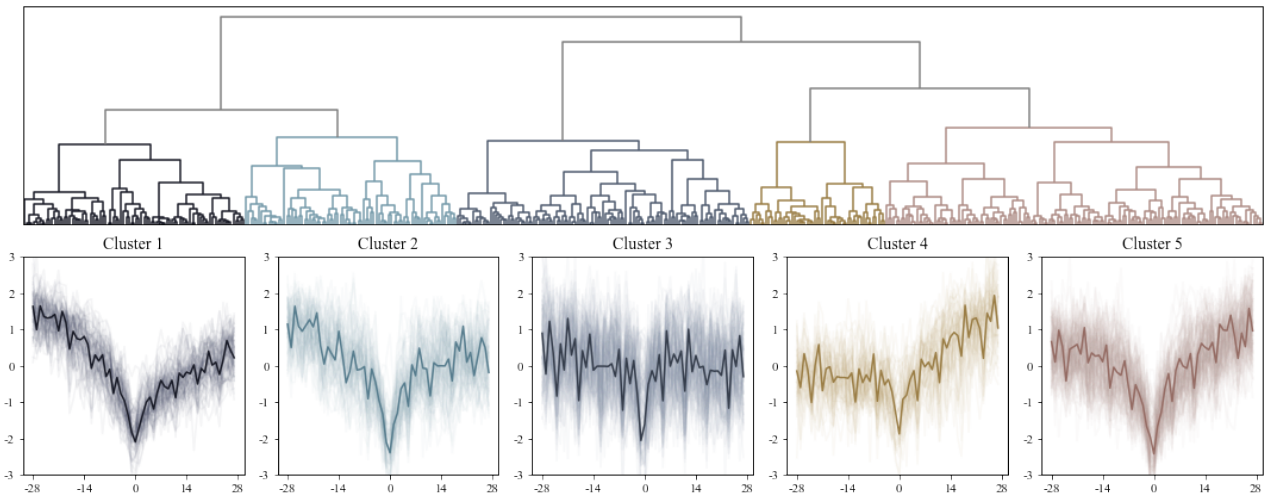

Figure 1: Data-driven event typology from a Dutch newspaper data set (1950-90) based on the *Embed, Detect and Describe* framework [8]. First, we cluster event flows based on surprise for newspapers (upper) to identify archetypal event signatures (lower). Bold lines (lower) indicate the archetypal event flow modeled with dynamic-time warping barycenter averaging from the underlying empirical event flows in thin lines. The archetypes event flows capture how events impacted the news in five characteristic manners and, by extension, how events impacted our historical temporality.

References

[1] Theo Jung and Anna Karla. 1. Times of the Event: An Introduction. *History and Theory*, 60(1):75–85, 2021.

[2] N. H. Packard, J. P. Crutchfield, J. D. Farmer, and R. S. Shaw. Geometry from a time series. *Phys. Rev. Lett.*, 45(9):712–716, 1980. Publisher: American Physical Society.

[3] Floris Takens. Detecting strange attractors in turbulence. In David Rand and Lai-Sang Young, editors, *Dynamical Systems and Turbulence, Warwick 1980*, pages 366–381. Springer Berlin Heidelberg, 1981.

[4] Tim Sauer, James A. Yorke, and Martin Casdagli. Embedology. *Journal of Statistical Physics*, 65(3):579–616, 1991.

[5] Edward Ott. *Chaos in Dynamical Systems*. Cambridge University Press, 2 edition, 2002.

[6] Jianbo Gao, Yinhe Cao, Wen-wen Tung, and Jing Hu. *Multiscale Analysis of Complex Time Series: Integration of Chaos and Random Fractal Theory, and Beyond*. Wiley-Interscience, 1 edition edition, 2007.

[7] Antske Fokkens, Marieke Van Erp, Piek Vossen, Sara Tonelli, Willem Robert Van Hage, Luciano Serafini, Rachele Sprugnoli, and Jesper Hoeksema. GAF: A grounded annotation framework for events. In *Workshop on Events: Definition, Detection, Coreference, and Representation*, pages 11–20, 2013.

[8] Melvin Wevers, Jan Kostkan, and Kristoffer L. Nielbo. Event Flow-How Events Shaped the Flow of the News. *Proceedings of the Second Computational Humanities Research Conference (CHR2021)*, 1613:0073, 2021.