Community detection in networks with unobserved edges

Till Hoffmann[1], Leto Peel[2] Renaud Lambiotte[3], and Nick S. Jones[1,4]

[1] Department of Mathematics, Imperial College, London SW7 2AZ, UK, tah13@imperial.ac.uk
[2] ICTEAM, Université catholique de Louvain, Louvain-la-Neuve B-1348, Belgium, leto.peel@uclouvain.be
[3] Mathematical Institute, University of Oxford, Oxford, OX2 6GG, UK, renaud.lambiotte@maths.ox.ac.uk
[4] EPSRC Centre for Mathematics of Precision Healthcare, Imperial College, London SW7 2AZ, UK, nick.jones@imperial.ac.uk

Detecting communities in networks provides a means of coarse-graining the complex interactions or relations (represented by network edges) between entities (represented by nodes) and offer a more interpretable summary of a complex system. However, in many complex systems the exact relationship between entities is either unknown or unobserved. Instead, we may observe interdependent signals from the nodes, such as time series, which we may use to infer these relationships. Over the past decade, a multitude of algorithms have been developed to group multivariate time series into communities with applications in finance, neuroscience, and climate research. For example, identifying communities of assets whose prices vary coherently can help investors gain a deeper understanding of the foreign exchange market [2, 3] or manage their market risk by investing in assets belonging to different communities [4]. Global factors affecting our climate are reflected in the community structure derived from sea surface temperatures [5]. Current methods for detecting communities when network edges are unobserved, typically involve a complicated process that is highly sensitive to specific design decisions and parameter choices. In this work, we develop a Bayesian hierarchical model for multivariate time series data that provides an end-to-end community detection algorithm that does not extract information as a sequence of point estimates, but instead propagates uncertainties from the raw data to the community labels.

The variability of high-dimensional time series is often the result of a small number of common, underlying factors [1]. For example, the stock price of oil and gas companies may be positively affected by rising oil prices, whereas the manufacturing industry, which consumes oil and gas, is likely to suffer from rising oil prices. Motivated by this observation, we model the multivariate time series y using a latent factor model, i.e. the n-dimesional observations at each time step t are generated by a linear transformation A of a lower-dimensional, latent time series x and additive observation noise. The entries A_{iq} of the $n \times p$ factor loading matrix encode how the observations of time series i are affected by the latent factor q. Using our earlier example, the entry of A connecting an oil company with the (unobserved) oil price would be positive, whereas the corresponding entry for an automobile company would be negative.

Our approach naturally supports multiscale community detection as well as the selection of an optimal scale using model comparison. We validate and study the prop-

erties of the algorithm using a series of synthetic datasets. We then apply it to daily returns of constituents of the S&P100 index to identify salient communities of similar stocks and to climate data of US cities to identify homogeneous climate zones. Figure 1 shows the detected communities from the daily returns of constituents of the S&P100 index of the stocks of 100 large companies in the United States. We obtained 252 daily closing prices for all stocks during 2016[5]. The community assignments capture salient structure in the data. For example, the three smallest communities each having only two members consist of two credit card companies, two defence companies, and two chemical companies (which have since merged to form the conglomerate DowDuPont). Other specialised communities consist of financial services companies (e.g., Citigroup, Goldman Sachs), as well as manufacturing and shipping (e.g., Boeing, Caterpillar, FedEx, United Parcel Service).

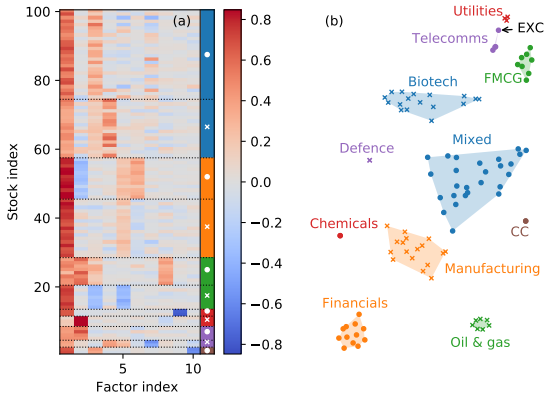

Fig. 1. Detected communities of stocks are correlated with industry sectors. (a) heat map of the factor loadings inferred. Each row corresponds to a stock and each column corresponds to a factor. (b) a two-dimensional t-SNE embedding of the factor loading matrix with cluster labels.

References

1. Fama, E.F., French, K.R.: Common risk factors in the returns on stocks and bonds. Journal of Financial Economics 33(1), 3–56 (1993)
2. Fenn, D.J., Porter, M.A., McDonald, M., Williams, S., Johnson, N.F., Jones, N.S.: Dynamic communities in multichannel data: An application to the foreign exchange market during the 2007–2008 credit crisis. Chaos 19(3), 033119 (2009)
3. Fenn, D.J., Porter, M.A., Mucha, P.J., McDonald, M., Williams, S., Johnson, N.F., Jones, N.S.: Dynamical clustering of exchange rates. Quantitative Finance 12(10), 1493–1520 (2012)
4. MacMahon, M., Garlaschelli, D.: Community detection for correlation matrices. Phys. Rev. X 5(2), 021006 (2015)
5. Tantet, A., Dijkstra, H.A.: An interaction network perspective on the relation between patterns of sea surface temperature variability and global mean surface temperature. Earth System Dynamics 5(1), 1–14 (2014)

[5] https://finance.yahoo.com/