

International Journal of Population Data Science

Journal Website: www.ijpds.org



Infrastructure for analyzing population scale network embeddings

Tom Emery¹

¹ODISSEI, Rotterdam, Netherlands

Nation-scale social networks derived from administrative registers create new possibilities for longitudinal social science, but they also impose substantial infrastructure demands. This paper documents the infrastructural requirements underlying the construction and use of node embeddings for a dynamic, multiplex population network covering the entire population of the Netherlands. The network consists of roughly 17 million individuals connected by more than one billion typed relations across family, household, neighborhood, workplace, and education layers, with annual snapshots available since 2009. The analysis requires secure access to highly sensitive administrative microdata, scalable storage for multi-year billion-edge graphs, and high-performance compute capable of executing random-walk-based embedding algorithms at population scale. Temporal analysis introduces further demands: embeddings must be aligned across years into a common latent space without leaking future information, requiring repeated large-scale linear transformations over tens of millions of vectors. Downstream use of embeddings amplifies these requirements. Integrating network-derived representations with administrative and survey data for prediction tasks depends on reliable data linkage, consistent identifiers, and reproducible pipelines. Transforming continuous embeddings into balanced, reusable features through whitening and Fibonacci-based equipartitioning requires memory-efficient linear algebra and stable execution environments. By foregrounding these infrastructural dependencies, the presentation demonstrates that population-scale embedding analysis depends on tightly integrated systems for secure data access, storage, computation, and reproducibility. The contribution is therefore not only methodological, but architectural: it provides a concrete illustration of the infrastructure stack required to turn whole-population networks into longitudinal, reusable research assets for the social sciences.

