Case for Support

Introduction

This project will join together social scientists with researchers based in the formal and physical sciences (physics, computer science, mathematics, and statistics) to develop generative latent variable models for networks capable of addressing some fundamental challenges of social network datasets. Namely, we will address the problems of treating noise in self-report nominations, the integration of various multi-level units, and the time dynamics of network structure. While these models will be of use for many social scientists, they will be of immediate use for a collaborative, cross-cultural comparative project studying social and economic inequality in over forty communities around the globe.

Since 2009, when Borgatti and colleagues heralded the florescence of network analysis in the social sciences (Borgatti et al., 2009), social network analysis has only continued to flourish. Network data sets gleaned from surveys (Banerjee et al., 2013), social media (Park et al., 2018), official records (DellaPosta, 2017), and geospatial trackers (Migliano et al., 2017) can now be found in journals spanning sociology (Hofstra et al., 2017), anthropology (Power and Ready, 2018), economics (Alatas et al., 2016), psychology (van Zalk et al., 2019), demography (Helleringer et al., 2009), geography (Ter Wal, 2013), political science (Leifeld and Schneider, 2012), and medicine (Kim et al., 2015). Common across these studies is a recognition of the importance of capturing the full richness of the social context, often by gathering information on multiple types of social connections and by recording extensive details on individual attributes.

Despite this richness, the methods used to analyse these complex and multifaceted datasets often entail substantial reductions and simplifications. Much work continues to be descriptive in nature, presenting only basic summary statistics of network structure or of individuals' positions within it. Attempts at inference often continue to rely on relatively simple tools, with techniques such as quadratic assignment procedure, exponential random graph models, and stochastic actor-oriented models being seen as the state-of-the-art (Ward et al., 2011; Cranmer et al., 2017; Snijders, 2010; Glückler, 2007; Glückler and Doreian, 2016; Ter Wal and Boschma, 2009; Robins, 2013; Lubell et al., 2012; Perkins et al., 2015; Sweet, 2016). Whether descriptive or inferential, many of the commonly used techniques require frustrating simplification of the data. Directed relationships are treated as if they were symmetric. Multiple distinct relationship types are either combined to create a single binary link or studied in isolation. Reports from multiple individuals are combined to represent a collective unit such as a household.

Often, these simplifications are made because the analytical tools that social scientists use cannot account for these complexities, working only with single matrices or static sets of individuals. There remains, therefore, a need for greater dialogue between the social and physical sciences (as Borgatti et al. (2009) called for) to facilitate the development and dissemination of tools that do not require such simplification. We highlight three features of social network data that should be considered in such models.

First, people's representations of their social worlds are not impartial accounts of an objective reality (Krackhardt, 1987; Freeman, 1992; Brashears, 2013). There are likely to be substantial biases and inaccuracies in networks constructed on the basis of individual self-reports. Indeed, there is a rich literature documenting issues of informant accuracy (Killworth and Bernard, 1976; Bernard et al., 1984), with evidence that recall over even short periods of time can be low (Adams et al., 2006; Bell et al., 2007), that certain types of alters and relationships are more readily named (Marin, 2004; Shakya et al., 2017; Marineau et al., 2018), and that the order of questions and mode of elicitation can influence responses (Pustejovsky and Spillane, 2009; Eagle and Proeschold-Bell, 2015). Consequently, researchers are developing techniques to model these biases (Comola and Fafchamps, 2017), with Bayesian approaches (Butts, 2003) and latent network models (Hoff et al., 2002) holding particular promise.

Second, people are often members of larger social groupings, such as families or firms, that form hierarchically nested groups (Zhou et al., 2005; Hamilton et al., 2007; Hill et al., 2008). Because of this, individuals are sometimes asked to speak on behalf of a higher-level unit, as when a "head of household" answers questions for the entire household. Rather than ignoring such nestedness or studying different levels in isolation, new techniques are being developed to allow for their simultaneous modelling (Lazega et al., 2008; Lazega and Snijders, 2015; Koster et al., 2015; Koster, 2018; Montiglio et al., 2020; Wang et al., 2013, 2016).

Third, people's relationships are in constant flux, as are many of their attributes (e.g., their education, wealth, or status) and their group memberships (e.g., their household or company of work). Studying such changes (and the direction of their temporal associations) is crucial for establishing causality. These

changes can, however, raise substantial issues for modelling dynamic networks. For example, when a household (or any higher-level unit) can change composition over time, the fundamental unit of analysis may be impermanent. While this changeable nature has been tackled in other areas of social statistics (Steele, 2008; Steele et al., 2013, 2019), it has not been resolved for networks.

Building on these insights, we will develop a class of generative network models that appropriately model the complexity of social networks. Our novel framework will provide a principled approach that can account for the characteristics of the network data typically collected by social scientists, namely the noisiness of self-report nominations and the nestedness of individuals into higher-level units. To do this, we have recruited a small team of researchers from a wide range of disciplines—mathematics (Bianconi), statistics (Sweet, Steele), computer science (Sridhar, Valera), physics (De Bacco), anthropology (Power, Koster, McElreath), and psychology (Redhead)—facilitating the dialogue that Borgatti et al. (2009) called for.

We are grounded in this effort by the needs of a particular project: a cross-cultural, comparative and longitudinal study of social and economic inequality, led by two of the co-investigators. Called by the acronym "ENDOW" (**E**conomic **N**etworks and the **D**ynamics **O**f **W**ealth Inequality) and funded by the US National Science Foundation (`https://www.nsf.gov/awardsearch/showAward?AWD_ID=1743019`), this project has enlisted anthropologists working in over thirty countries around the world to gather comparable social network data in over forty communities (`https://endowproject.github.io`). With this project, we have all of the crucial data complexities discussed above: network data reported by individuals, grouped into households (variably defined across communities and variably stable across time), based on multiple name generators, gathered at multiple time points. There is immense potential in this project to answer core questions about the drivers of inequality around the globe, but such comparative work needs to be done with care and consideration. For example, a "household" may mean very different things in different sociocultural settings (Randall et al., 2011; Randall and Coast, 2015), complicating the easy comparison of household-level networks between communities. The need to develop analytical tools that can appropriately model these rich data has been the impetus for this proposal.

Research Questions

Grounded in the specific research aims of the ENDOW project, and cognisant of the generality of these measurement and modelling issues, we ask:

1. How do we account for the **individual-level biases** in reporting relationships that are introduced when constructing networks from self-report data?

2. How do we model the **multi-level nature** of many social networks, where individuals can be nested in higher-level units (e.g., households or companies) that can vary in composition across contexts?

3. Building upon the static formulations above, how can we model the **time evolution** of such biased, multi-level networks?

Modelling Approach

In this project, we will develop a general and flexible generative network modelling framework. We consider that of probabilistic latent variable models our main approach. These are powerful tools for modelling complex distributions of random variables, e.g., systems with many individual variables interacting as in a social network; under certain conditions (that often apply to the systems studied here), their validity is guaranteed by formal mathematical theorems (De Finetti, 1929; Hoover, 1982; Aldous, 1985). They can capture complex dependencies between data and allow for more straightforward parameter interpretation (Hoff, 2008). Among latent variable models, we consider variants of latent class models (or stochastic blockmodels) (Holland et al., 1983). These allow for a relevant interpretation of the latent variables as community membership: an individual can be assigned to communities (or groups) based on the interaction pattern that is observed from data. In particular, we will consider overlapping memberships, where individuals can belong to multiple communities, as this is more realistic for the scenarios considered here. Our foundational model will be grounded in fundamental observations about the characteristics and generation of social network datasets, most notably their noisiness and their nestedness, and be purposefully general to allow for future extensions. The input and guidance from the proposed research team is essential for us to establish the details of this approach. By drawing on their

complementary expertise, this project provides a unique opportunity for the basic framework outlined here to be advanced.

Individual-level biases: accounting for noisy data and double-sampled questions Most social network analysis assumes that self-report data provide a complete and accurate representation of the social network. This is, however, too strict an assumption when considering that people are prone to making errors and may harbour biases when reporting on their social relationships. As a result, these self-reported data may specify the network structure imperfectly and lead to fundamentally wrong estimates of network properties and to misleading conclusions (Kossinets, 2006). To partially address this issue, a common technique is to "double sample" questions. For example, ENDOW project surveys asked respondents not only to name who they went to for different types of help, but also who came to them. When combined with complete sampling, this can give two perspectives on what should be the same relationship. While this provides important additional information, it does not resolve the potential biases and myopia in people's recall of their connections, taking people's perception of a social relation for the relation itself.

Related work: Previous probabilistic modelling approaches have attempted to account for network error and reconstruction. Here, we briefly overview probabilistic models as these are more flexible, e.g., they can perform tasks such as sampling new data or making predictions (though other approaches that aim at correcting errors have also been developed to address this problem). Notably, hierarchical Bayesian methods have been developed to estimate both the reliability of respondents when completing social network surveys and the structure of the social network in the presence of errors (Butts, 2003). Alongside this, two recent approaches have used the stochastic block model to obtain probabilistic estimates of network structure. The first proposes a link reliability measure to identify missing and spurious interactions in network observations (Guimerà and Sales-Pardo, 2009). The second splits the model into two terms, where the first term models the network structure and the second models the measurement process conditioned on the generated network (Newman, 2018; Peixoto, 2018). While these approaches have provided much-needed contributions, their applications are limited for several reasons. They do not explicitly consider information from double-sampled ties. Instead, they assume knowledge of the true underlying structure and consider synthetic random removal or addition of links to test the model on real networks or assume that edges of a network are measured directly and repeatedly. In addition, these methods lack the flexibility to model various types of information, such as multilayer, annotated, and weighted networks.

Proposed methodology: We will develop a flexible generative network model that explicitly accounts for "double sampled" data: for each possible existing tie between individual i and j, we can observe two noisy observations, one reported by i and one by j. The idea is to then treat the true underlying and unobserved tie as a latent variable itself, while the data to be fitted are the double-sampled ties. We aim to simultaneously infer both the structure of the network from noisy measurements and its division into communities. This, in turn, will aid us in estimating missing links and examining the reported links. We will use this model as a building block for subsequent extensions by integrating the two types of structure described below, namely multi-level and dynamical.

Multi-level nature: adding nestedness Social networks are intrinsically multi-level in nature. People may, for example, be grouped into households, companies, or schools. In the majority of empirical studies, networks have at least two distinct hierarchical levels, and there is often attribute information associated with each level. Importantly, ties may form between individuals from different higher-level units, relationships or exchanges may simultaneously be observed between the higher-level units, and there may also be interactions between individuals and higher-level units.

Related work: There are only a few prevailing latent variable hierarchical models on networks (Clauset et al., 2008; Peixoto, 2014; Sweet et al., 2013). While they provide an intuitive mathematical formalism for incorporating multi-level structures into the framework of latent class models, they lack flexibility as they make *a priori* assumptions of a specific hierarchical structure. Moreover, they do not allow for individuals to belong to more than one community and are valid only for single-layer networks.

Proposed methodology: We will develop a multilevel latent variable model that investigates how higher-level structures emerge from individuals. Specifically, we will model interactions both across and within levels, so that an individual i might interact with an individual j, but at the same time i indirectly interacts also with the other members of the higher-level unit that j is nested within. We intend to assign household-level latent variables that govern how individuals nested in higher-level units interact with individuals who are assigned to different higher-level units. In practice, these latent variables function as

high-level templates for the individual-level community membership, which are drawn as a perturbation of these high-level templates. In other words, explicit ties are observed between individuals, which depend on their individual community memberships. This in turn, depends on higher-level latent variables, thus creating an implicit dependence structure between all individuals assigned to a given higher-level unit with all individuals assigned to another higher-level unit, even though there may be no explicit ties that directly link them. The higher-level latent variables can then be used to assess observable quantities at this higher level, such as a household's material wealth. For instance, they can be used as input features to regression setups where the outcome variable is some measure of wealth. Note that, given this hierarchical latent structure, the proposed method is conceptually novel and provides a fruitful alternative to conventional approaches, which typically treat a household as the sum of the individuals within it.

Time evolution: adding dynamical network structure The units within our study systems change over time. This is valid for all hierarchical levels. For example, individuals may marry into an existing household, may start up a new household, or may leave the community. As a consequence, new households can emerge and old households can change composition or dissolve entirely. Across many study systems, we can observe both lower- and higher-level units shifting composition across time, a process more complex than a simpler reshuffling of individuals. The most promising avenue for capturing these dynamics is to model them using a dynamical community detection model.

Related work: Dynamical community detection is based on modelling temporal networks where nodes and links change over time. A standard approach is to use Markov Chains to model a dynamical stochastic block model (Peixoto and Rosvall, 2017; Matias and Miele, 2017; Fu et al., 2009). Another approach considers a temporal network as a series of snapshots, and thus models it as a multilayer network, with layers representing different time steps (Zhang et al., 2017; Ghasemian et al., 2016). To prevent identifiability issues (Matias and Miele, 2017), these models typically assume unchanging group memberships but correlated edges (Zhang et al., 2017) or, the converse, time-varying group memberships but independent edges between snapshots (Ghasemian et al., 2016). None of these methods have addressed the question of how dynamics impact lower and higher hierarchical levels. In addition, existing methods typically model either nodes or edges changing in time, while a multitude of empirical applications require both to be modelled.

Proposed methodology: We propose to develop a model that assumes a hierarchical block structure while allowing for the time evolution of both nodes and ties. In other words, we will integrate the previous two methods by adding the modelling of quantities changing in time. In particular, we will model nodes' community membership changing in time, while keeping the interactions between groups stable, as this seems the most plausible hypothesis for many systems studied by social and behavioural scientists. At the same time, we will model change in higher-level variables over time as a result of both individuals and ties changing their status. Our proposal to model dynamical membership latent variables at each hierarchical level, while capturing complex within- and between-level interactions, is a fundamentally novel contribution. By developing this framework we are taking a necessary and important step towards fully capturing time-evolving complex systems of this type. This, in turn, will allow the analysis of community structure at each hierarchical level across time and relate it to the concurrent time evolution of higher-level variables.

Applications & Impact

The ENDOW project is aimed at investigating the economic consequences of social network structure (Jackson et al., 2017), both for individuals and for the larger communities they comprise. This is a fundamentally comparative project, as we expect that the variation we observe in the structure of social networks will help to explain some of the cross-cultural variation in wealth inequality. A household's ability to increase its material wealth over time should be influenced by its position within the social network and the overall structure of that network (Kets et al., 2011). To evaluate these hypotheses, ENDOW team members have gathered records of households' material assets and household members' demographics, employment, educational attainment, etc. that can be combined with responses to social support questions to generate multilayer networks with rich metadata.

Analysing these networks, however, is complicated both by variation in measurement across communities (e.g., differences in survey questions and survey respondents) and also by variation in household composition (both across communities and across time). The techniques that we will develop in the proposed project provide novel solutions to these knotty issues, allowing us to answer the core ques-

tions of this substantial research project. Given the growth of inequality (whether of income, wealth, opportunity, living conditions, etc.) in the last four decades (Alvaredo et al., 2018), a more dynamic understanding of how inequality emerges is a crucial task for global society.

Importantly, the ENDOW project introduces a large number of social scientists to the collection and analysis of social network data. Beyond the comparative projects that will come out of this collaboration, there are the many independent projects that ENDOW contributors will develop using their own network data. We therefore have an important opportunity (and responsibility) to ensure that the 40-plus ENDOW researchers analyse their network data in appropriate, principled ways. Thanks to the disciplinary range of the ENDOW team (working in anthropology, economics, international development, public health, etc.), we expect our models to be used in journal articles spanning the full range of the social sciences, as well as in white papers and policy briefs in the areas of health, development, and conservation.

While the value for the ENDOW project is clear, we see a wide breadth of potential applications of these models. As outlined at the start of this proposal, there is a growing interest in networks across the social sciences coupled with a growing supply of complex network datasets, many of which have the same features as our ENDOW network data. Take, for example, the paper by Banerjee et al. (2018), drawing on two rounds of network data from 75 villages in India. The authors combine all network layers into a single unweighted, undirected network with nodes representing households, and they skirt the issue of changes in household composition across time by removing those households that were not present at both time points. Decisions like these are understandable, given the tools that social scientists have at their disposal. But researchers in a variety of fields are increasingly dissatisfied with such simplifications, calling for a greater attentiveness to the multilevel nature of networks (Lazega and Snijders, 2015) and to the imprecision with which they are measured (Ward et al., 2011). We will offer novel methods that will help social scientists heed such calls.

To ensure that our work has the desired impact, we will aim to publish in outlets that are widely read by potential practitioners, such as broad interdisciplinary journals (e.g., *Science Advances*, *Scientific Reports*, *Nature Communication*). We additionally plan publications introducing how the models can be used in practice, aimed at journals focused on networks or disciplinary methods journals (e.g., *Social Networks*, *Network Science*, *Sociological Methods & Research*, *Journal of Statistical Software*). All publications will be accompanied by open source code and eventually R and Python packages to facilitate uptake.

Timetable & Outputs

To develop these models, we will build on the expertise of the co-investigators and a small number of collaborators who span the social, physical, and formal sciences. The full team of researchers will convene at three workshops: a virtual one in January 2021, one held at the London School of Economics in June 2021, and one held at the Max Planck Institute for Evolutionary Anthropology in January 2022. We expect much of the time in these workshops to be spent in front of whiteboards, hashing out the details of the models. In addition, the co-investigators will meet twice (once at the Networks 2021 conference in Washington, DC in July 2021 and again at the Max Planck Institute for Intelligent Systems in February 2022) to focus on finalising the models and code and drafting articles.

We have already invited and received confirmation for the following researchers: **Ginestra Bianconi**, Professor in Applied Mathematics at Queen Mary University of London, is a mathematician working on complex networks, with a particular focus on multilayer networks (e.g., Boccaletti et al., 2014; Bianconi, 2018). **Jeremy Koster**, Associate Professor in the Department of Anthropology at the University of Cincinnati, is an anthropologist who studies social networks in Nicaragua, developing methods for their study (e.g., Koster and Leckie, 2014; Koster et al., 2015), and is the co-director of the ENDOW project. **Richard McElreath**, Director of the Department of Human Behavior, Ecology, and Culture at the Max Planck Institute for Evolutionary Anthropology, is an anthropologist and statistician developing Bayesian statistical tools in R for social scientists (e.g., McElreath, 2016). **Dhanya Sridhar**, postdoctoral researcher at the Data Science Institute at Columbia University, is a computer scientist working on applied causal inference (e.g., Sridhar and Getoor, 2019; Wang et al., 2019). **Tracy Sweet**, Associate Professor in the Department of Human Development and Quantitative Methodology at the University of Maryland, is a statistician developing multilevel network models, often for educational contexts (e.g., Sweet et al., 2013, 2019). **Isabel Valera**, Professor at Saarland University, is an engineer and computer scientist who works on machine learning and probabilistic inference with interdisciplinary applications (e.g., Valera et al., 2017; Kilbertus et al., 2019; Vergari et al., 2019). Additionally, **Fiona Steele**, Pro-

fessor in Statistics at the London School of Economics, who has developed longitudinal multilevel models allowing for evolving household structure (e.g. Steele et al., 2013, 2019), will serve in an advisory capacity. With this set of experts, we are confident that the models we develop will be of great value not only to the ENDOW project, but to a wide range of applications in fields as varied as sociology, demography, education, and computer science.

The primary outputs of this collaborative project will be articles submitted to peer-reviewed journals and the Python and R code, to be wrapped into packages to facilitate their use. The planned publications build sequentially to offer increasingly complex network models.

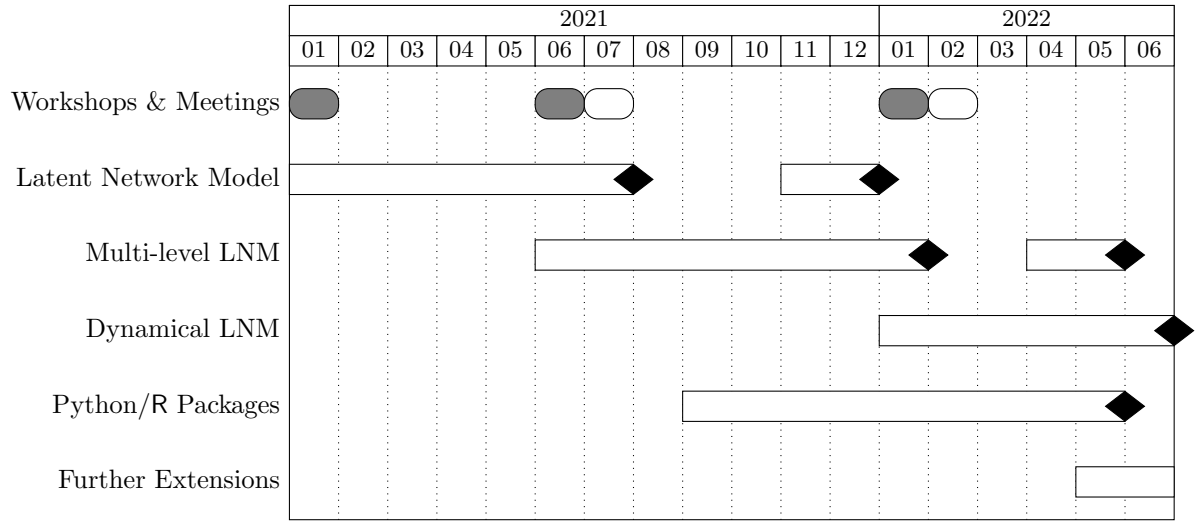
First, we will develop a foundational latent network model that grapples with the various biases inherent to self-report network data, to be discussed extensively at the virtual workshop in January 2021. We expect to post an open-access preprint (likely on the ArXiv), make R and/or Python code publicly available on GitHub, and submit this initial article to a peer-reviewed journal soon after our meetings in the summer of 2021, with revisions and final publication happening within the time frame of this grant.

Second, we will extend the latent network model to account for the hierarchical nestedness of much social network data. We plan to start sketching out this model at the first in-person workshop in London in June 2021, leading to a draft to be discussed at the workshop in January 2022. We expect to be able to post a preprint and code and submit for publication soon after that workshop, for final publication before the end of the grant.

Third, we will transition from the static latent network model to a dynamic model with continuous time. This will be a main focus of the workshop in Leipzig in January 2022. We expect to be able to have an initial preprint posted (along with a code repository) before the end of the grant, for eventual submission to a journal.

Finally, we place great importance on making these models accessible to as many social scientists as possible, and so want to develop these models into R and Python packages centred on latent network models. We will therefore hire a research officer (for a nine month appointment, September 2021 to May 2022) to package our code. Alongside the package, we will write a paper outlining its use with worked examples. We expect to have packages released and a preprint of this tutorial posted before the end of the grant.

Importantly, we have plans to build on these models further and see this current set of outputs as the start of an ongoing research agenda and collaboration.



(Diamonds are outputs (preprint or journal article); grey ovals are workshops; white ovals are co-investigator meetings).