Internal migrations in Austria: modeling and inference of temporal graphs

Thomas Robiglio

Dept. of Network and Data Science, Central European University, Vienna, Austria

Supervisors: Tiago P. Peixoto and Márton Karsai

robiglio_thomas@phd.ceu.edu

May 24, 2024

Abstract

Migration plays a crucial role in socioeconomic development and is related to many phenomena relevant to today's society. Understanding the driving forces behind these mobility phenomena can provide insights for better policy-making. Migration events are intrinsically described as temporal, annotated, and directed graphs. Thus a data-driven analysis of these processes calls for the employment of tools from network science. Register-based data about internal relocations in Austria over the last twenty years from Austria's Federal Statistical Office allows for a first-in-kind large-scale network analysis of internal migration phenomena in Austria. This PhD project aims to develop an inferential framework for studying networks of migration flows. This will be achieved with an effective modeling approach, using generative models for temporal graphs, and a mechanistic modeling approach, using minimal models able to reproduce the patterns observed in the data and identify the key driving factor of migration phenomena. The combination of these two complementary approaches into a mixed effective-mechanistic model will contribute to building the methodological backbone of the analysis of the structure, driving forces, and implications of internal migrations in Austria.

I. Introduction

a. Main motivation: understanding internal migrations

Migration plays a central role in urbanization, segregation, gentrification, and in many phenomena related to socioeconomic development [1]. The driving forces of migration can be extremely varied, including labor market imbalances, wealth inequalities, conflicts, and ethno-racial segregation—reflecting the rapid increase in complexity of human societies [2]

For these reasons, migration is among the crucial topics in regional and national governance. This has led to significant attention to migration in research. Researchers from various fields investigate why people migrate, how migration takes place, and what the consequences of migration are in a broad sense, both for migrants themselves and for societies involved in migration [3].

While a lot of attention has been paid to international migration in applied research, many questions are still unanswered for a comprehensive understanding of internal migration. Data shows that the majority of migrations occurring in the world happen within national boundaries [4], thus narrowing the research focus on international migration neglects a large portion of the overall volume of the events under study. Finally, focusing on international migrations limits our ability to assess social dynamics and the impact of local policy.

b. Migrations and network science

A data-driven analysis of migration phenomena requires tools from network science [5, 6]. This is because phenomena of human mobility are fundamentally relational: different spatial regions (*e.g.* municipalities in a country) represent nodes of a network, and the movement of people from a source to a destination represents weighted, directed, and time-annotated links between such nodes (see Fig. 1).

The development of network science has been driven by the recent deluge of empirical data. In some cases, however, network methodology has outpaced data collection. This is precisely the case for migration and mobility [7], where data has been available only at a coarse-grained spatio-temporal level. Although the state-of-the-art network analysis methodology could provide in principle an all-encompassing, multimodal understanding of global migration flows, the available data confined it so far only to a limited overview.

c. Problems related to data quality

Large-scale scientific studies regarding migrations generally face multiple problems related to data quality [8]. These problems generally trace back to the low resolution of the data and to the fact that the data originates from different sources.

Historically, mostly aggregated and sampled data has been available for large-scale scientific studies. When investigating international migrations, researchers commonly examine data with yearly resolution and spatial details at the country level. For internal migrations, data is rarely available with a resolution below the level of municipalities. This lack of detailed data hinders researchers' capacity to analyze migration phenomena at a finer, more local level. For instance, in internal migrations, data limited to municipal resolution overlooks relocation dynamics within major cities. Such local phenomena could be linked to social processes of crucial importance like segregation or gentrification.

The second limitation in the data quality in large-scale studies regarding migration is due to the use of different sources for the empirical data. These different sources can have different sampling methodologies and even different definitions of what a migration event consists of. Such diversity in definitions and methodologies among data sources can result in data inconsistencies, potentially introducing biases in researchers' findings [9].

These data quality issues are significantly exacerbated when we attempt to connect patterns of migration to other socioeconomic factors, such as employment, income, age, sex, and many others. In these cases, we need to integrate even more heterogeneous data sources, implemented often for different purposes, with

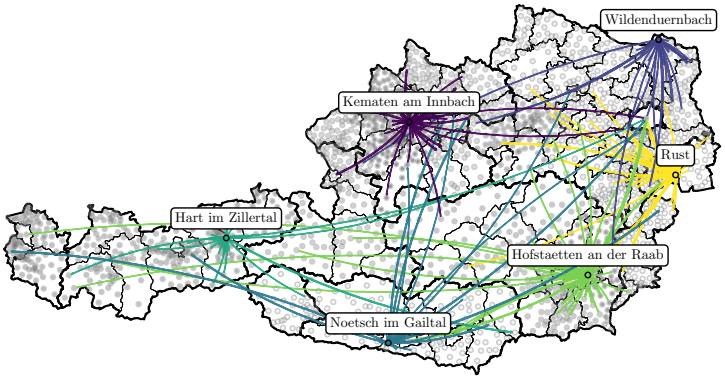

Figure 1: **Network of internal migrations in Austria.** Relocations in 2021 from and to six selected municipalities. The colored lines represent the directed flow of relocations between municipalities, with the thickness of the line being proportional to the magnitude of the flow. We show administrative boundaries between federal states (*bundesländer*, thicker lines) and between statutory cities and districts (*statutarstädte*, *bezirke*, thinner lines).

non-overlapping identifiers, temporal and demographic coverage, and having their own undersampling problems.

d. Relocation data from Austria, the MOMA project

All these obstacles can be overcome with the opportunity to use nationwide administrative data that has become recently available from Austria's Federal Statistical Office: Statistik Austria. This agency has been collecting individual-level information on migration, based on address registration, as well as information on income tax, employment, and several demographic parameters for a near-complete fraction of the Austrian population, as well as businesses and industry, for 20 years (2002-2022). This kind of register-based data has near-complete temporal and demographic coverage, therefore can essentially eliminate the aforementioned limitations that prevent a comprehensive understanding of migration dynamics. Furthermore, the individual datasets can be cross-indexed, yielding an unprecedented window into the detailed connections between demographic dynamics and socioeconomic development, and a unique opportunity to identify their main driving factors.

This PhD project is part of the "Multiscale network modeling of migration flows in Austria (MOMA)" project[1] founded by WWTF (Vienna Science and Technology Fund). This project aims to reveal the nation-wide, multi-scale, hierarchical internal flow of people in Austria over two decades, together with its latent social and economic correlates. No similar research has ever been done for the population in Austria (and rarely in any other country), where similar population-scale registry datasets have become available for research only recently.

e. Modeling and inference of temporal graphs

Using such fine-grained data and the straightforward network representation of migration flows requires the development of appropriate models and algorithmically efficient methods. Unlike traditional spatial or time-series data, networks are sparse high-dimensional objects, and their analysis requires theoretical frameworks that go well beyond conventional methods.

The objective of this Ph.D. project is to establish a methodological framework for examining networks of migration flows. Two complementary modeling approaches will be used:

- *Effective modeling*, through the development and inference of generative models for temporal graphs.

[1] https://www.wwtf.at/funding/programmes/ess/ESS22-032/

- *Mechanistic modeling*, using minimal processes to qualitatively reproduce the patterns in the data and identify the relevant forces driving them.

Combining these two descriptions into a class of *mixed effective-mechanistic models* to analyze internal migrations in Austria will contribute to a large-scale understanding of these phenomena's driving forces and socioeconomic implications.

Beyond the specific goal of modeling internal migrations in Austria, this project will provide important methodological contributions to the network scientific field. Firstly, the development of new generative models for temporal graphs could be useful to study different systems where the temporal evolution of interactions is relevant, ranging from social dynamics, epidemics, and brain activity [10]. Secondly, the development of the mixed effective-mechanistic model will advance the literature on disentangling confounding network formation mechanisms [11]. Overall, the project will be carried out within the framework of inferential network science, contributing to the development of robust methodological tools for network data analysis in different fields of application [12].

II. Data

We will use the service "Microdata For Research" provided by Statistik Austria (Austria's federal statistical office) and the Austrian Micro Data Center (AMDC). This service enables access to individual-level administrative data on migration, based on address registration, together with data on income tax declaration, demographics, service, and trade.

a. Data description

The main dataset we will use in our analysis is `MIGSTAT - Wanderungsstatistik`. This dataset contains all relocations of the Austrian population, associated with a change of main residence under registration law, for the period 2002-2022. The migration statistics provide information on external migrations, *i.e.* immigrations from abroad to Austria, ($\sim 1-2\times10^5$ per year) and departures from Austria ($\sim 7-10\times10^4$ per year), as well as on internal migrations, *i.e.* change of residence between and within Austrian municipalities ($\sim 6.5-8\times10^5$ per year).

This dataset contains the records of all internal and external migration events in Austria for 20 years. It also includes spatial, temporal, and demographic variables such as nationality, gender, exact date of relocation, and address of origin and destination.

In the development of our analysis, we will also use a publicly available coarse-grained version of the dataset. The coarse-grained dataset has a spatial resolution limited to municipalities (except the city of Vienna where data is available at the level of districts) and yearly temporal resolution. This version of the dataset can be recovered from Statistik Austria's website[2] and easily accessed through the Netzschleuder network repository [13].

b. Data security and management

The full-resolution dataset is managed by the Austrian Micro Data Center (AMDC) and will be accessed remotely through the services provided by AMDC. The researchers involved in the project will abide by the technical and organizational measures to ensure and maintain data security[3] set by the AMDC.

III. Main contributions

When modeling human mobility researchers typically follow two approaches, we refer to these two approaches as *effective* and *mechanistic* modeling. On the one hand, effective modeling is concerned with quantitatively describing the statistical properties of the patterns observed in the data. On the other

[2] https://data.statistik.gv.at/web/meta.jsp?dataset=OGDEXT_BINNENWAND_1
[3] https://www.statistik.at/fileadmin/pages/1805/TOMs_AMDC_public_en.pdf

hand, mechanistic modeling aims at identifying the driving forces behind mobility phenomena. This is achieved with simple toy models encoding the hypothesis to be tested. Below, we present the role that these two modeling approaches will have in our large-scale analysis of internal migrations as well as the novel contributions that will be made in that context.

Effective and mechanistic modeling approaches are typically pursued as completely disconnected approaches. In the third subsection, we present a novel approach to modeling mobility phenomena through a class of mixed effective-mechanistic models combining these two approaches.

a. Effective modeling: generative models and inference of temporal graphs

Our data-driven analysis will be based on principled methods from statistical inference [12]. The general setting of this approach consists of considering the observed data $\mathcal{D}$ as the outcome of a—possibly uncertain—measurement process of a true unobserved network $\mathbf{A}$. This measurement step is modeled as a generative process $P(\mathcal{D}|\mathbf{A})$, consequently estimates regarding the object of our study—the unobserved network $\mathbf{A}$—can be accessed through the posterior distribution $P(\mathbf{A}|\mathcal{D})$ computed with Bayes' theorem.

In this inferential framework, we can achieve an effective description of the migration phenomena in two ways. If we are interested in a system-level analysis, we can employ the commonly used structural network descriptors (*e.g.* clustering coefficient, density, *etc.*). In this case, sampling from the posterior distribution $P(\mathbf{A}|\mathcal{D})$ allows us to fully characterize these quantities and their uncertainties. Otherwise, we can also achieve a mesoscopic description of our system through the use of *generative models* encoding such a mesoscopic description. This consists of defining a model $P(\mathbf{A}|\{\theta\})$ for our network, where the parameters $\{\theta\}$ encode the mesoscopic description (*e.g.* the communities) we want to recover.[4]

The flow of internal migrations in Austria that is the object of our study is inherently described as a *weighted*, *directed*, and *temporal* network. We will employ existing generative models for temporal networks but also define new generative ones. These new models will be useful for our specific scope of modeling internal migrations in Austria but could be employed in other settings as well.

In the context of migrations, analyzing the patterns uncovered by these models will help us understand how flows of relocations change over time, what impact they have on socioeconomic development, how they are associated with gradual social processes such as segregation and urbanization or, furthermore, affected by policies or sudden shocks such as lockdowns during the COVID-19 pandemic or the 2015 European migrant crisis.

The modeling of temporal graphs usually takes the form of a set of snapshots of the network structure at different times [10]. This means that the fundamental unit of analysis is not a single network but the entire *history* of a network. Thus, the appropriate models for temporal graphs should generate entire histories, in the form of layered or temporally annotated adjacency matrices $\{\mathbf{A}^t\}$.

The existing generative models for temporal graphs can generally be grouped into two categories (Fig. 2). Approaches belonging to the first category model temporal correlations between interactions using Markov chains with short-term memory. Such approaches are based on the fact that edges appear and disappear by making transitions from present to absent and *vice versa* based on the control parameters and the assumptions of the different models. This category of models includes models based on the activity of nodes or edges [14, 15], and generalizations of static network models using discrete or continuous time Markov processes [16, 17]. Such models can capture structural features of real-world networks such as scale-freeness or—most notably—the presence of large-scale modular structures but rely on fixed rules governing the temporal dynamic, thus allowing for a description limited to stationary regimes. The second category of models focuses on the dynamic of networks at longer time scales, using network snapshots corresponding to different regimes of the dynamical process underlying the network and change points when the dynamic changes [18, 19]. The goal of this second class of models is thus to aggregate the structural dynamic of the system focusing on its long-term evolution.

[4] It is important to understand that these two sides of the effective description of a networked system are not distinct. Defining the generative process $P(\mathcal{D}|\mathbf{A})$ requires having a model for $\mathbf{A}$, *i.e.* a prior about the network structure, that is encoded into a generative model.

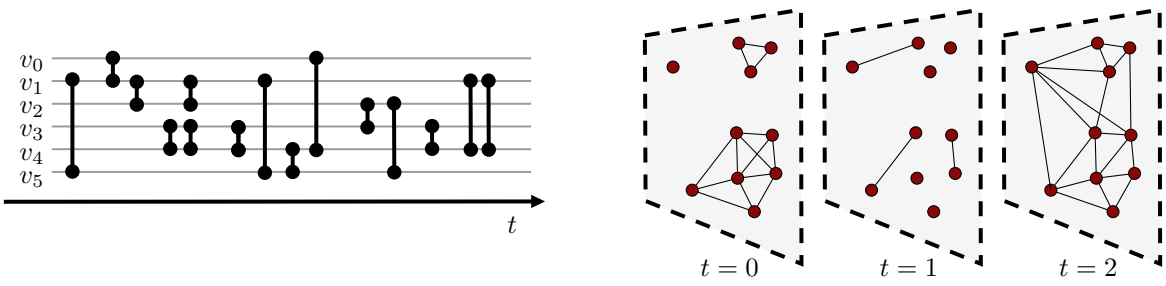

Figure 2: **The two approaches in modeling temporal graphs.** The first approach relies on short-time temporal scales, the presence/absence of edges is modeled as Markov chains (left). The second approach models longer time scales, this is done by aggregating the evolution of the network into different layers capturing the different regimes of the dynamic of the system (right).

Focusing on one timescale blurs the other, while in reality, many systems exhibit dynamics in a wide range of timescales. A full dynamical description of large-scale modular structures that combine the short-term evolution of community structure and change points controlling a longer timescale would further our understanding of the structure and dynamics of temporal networks [20].

In the existing generative models temporal networks defined as Markov chains, the communities and the change points work synergistically with the edge-level evolution. Building on these existing works we will formulate new generative models for temporal graphs with dynamic community structure and change points. The Bayesian formulations will protect us from overfitting, requiring more complex models only when they provide better compression (*i.e.* a better description) of the patterns observed in the data.

Below we will sketch two preliminary ideas to implement generative models with dynamic community structure and change points.

Stochastic block model with dynamic group assignments. This model is defined by considering the different snapshots of the network history $\{\mathbf{A}^t\}$ as degree-corrected stochastic block models (DC-SBM).[5] The dynamic of the network over time is encoded in the way we allow the usual parameters of the DC-SBM (group assignments, degree sequence, and group preferences) [21] to vary in time. The group preferences $\mathbf{e} = \{e_{rs}\}$ will be fixed in time, while we will have time-dependent group assignments $\{\mathbf{b}^t\} = \{\{b_i^t\}\}$ and degree sequences $\{\mathbf{k}^t\} = \{\{k_i^t\}\}$ (Fig. 3). Formulated this way, our model describes a temporal network with dynamic community structure, thus fulfilling parts of the goals outlined above. Moreover, this formulation does not require any kind of temporal aggregation of the observed network and could be used at the same temporal resolution of the data under study. In the case of temporally sparse data, where for many time points most of the nodes are inactive, our SBM will simply identify a partition of inactive nodes and the activation of a node will be represented as its transition to a partition of active nodes according to its connectivity pattern.

Stochastic block model with change points. A simple implementation of a model for temporal networks with multiple timescales is presented in Ref. [23]. Therein Peixoto and Gauvin propose a model in which the edges between N nodes of a network are placed dynamically according to a Markov chain with memory. The change points are introduced as points in time where the dynamic regime of the network changes abruptly. This is modeled in the Markov chain as points in time in which the transition matrix changes. The occurrence of such change points is governed by the probability q that one is inserted at any given time. With this model, from an observed time sequence of edges $\mathbf{s}$ we can recover (i) the order n of the Markov chain, (ii) the number of change points M, (iii) the probability of change points occurring q, (iv) the placements in time of the change points, and (v) the $M+1$ transition probability matrices $p^l_{x,\mathbf{x}}$

[5] In the following, we will often use simply SBM when referring to stochastic block models.

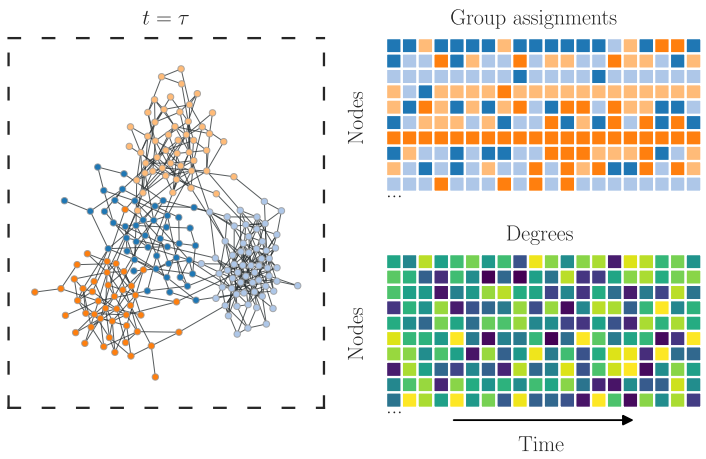

Figure 3: **Stochastic block model with dynamic group assignments.** In this model a layer at time $t = \tau$ is a sample from the microcanonical DC-SBM [22] ensemble with group assignments $\mathbf{b}^t$, degree sequence $\mathbf{k}^t$, and group preferences $\mathbf{e}$ (left). While the group preferences are fixed in time, the group assignments and the degree sequence are time-varying (right).

with $l \in [0, M]$. Inferring this full set of parameters $\{\theta\}$ from the posterior distribution $P(\{\theta\}|\mathbf{s})$ protects from overfitting, meaning that more complex models (many change points, long memory) are favored only when there is enough evidence in the data. In the original paper, the authors show the capacity of the model to reproduce the spreading of epidemics on the temporal network and capture through the change point relevant features of the data. For example, fitting the models to the `sp_highschool_new` dataset [24] (Fig. 4) separates between class hours (mostly localized interactions) and break times (more dense connections across partitions). In the original model by Peixoto and Gauvin, the edges are placed independently. Incorporating the community structure in this model can allow for a better compression of the observed networks. Such a model would provide a full dynamical description of large-scale modular structures that combines community structure and change points.

b. Mechanistic models of human mobility

Mechanistic models provide a complementary approach to the effective models described above. Models of this kind aim at reproducing the salient patterns observed in the data with a minimal set of processes and dynamical rules. Such processes and dynamical rules encode the driving phenomena behind the patterns in the data. This mechanistic modeling approach produces a qualitative rather than quantitative description of the phenomena under study.

In the context of human mobility, mechanistic models are widely employed. Prominent examples of models of this kind are gravity models [25], the radiation model of mobility [26], and Schelling's model of segregation [27].

Gravity models predicate that the frequency of interactions between two locations is proportional to the product of their population sizes and decays with the inverse of the distance between them. The idea behind this class of models is that migration is driven by the attractive force between source and destination countries and impeded by the costs of moving from one country to another. Such models were originally formulated to model the magnitude of stock trade between countries and later applied to human mobility. Generalized versions of this model include exponents for the quantities in play (population sizes and distance between locations) that deviate from unity and the integration of additional terms accounting for other demographic and socio-economic factors [28].

The radiation model was defined to describe job-seeking mobility. It predicates that individuals choose the closest job to their home, whose benefits are higher than the best offer available in their home location.

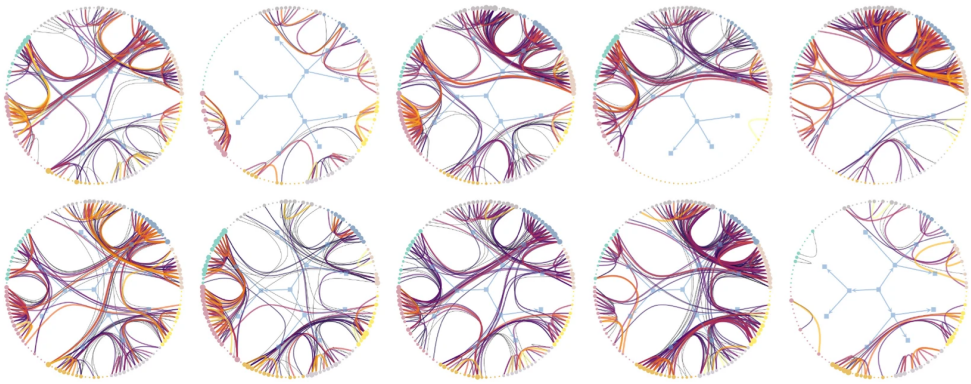

Figure 4: **Static Markov chain with change points.** Application of the model described in Ref. [23] to the `sp_highschool_new` dataset [24]. Network structure inside the first ten temporal segments, as captured by a layered hierarchical degree-corrected stochastic block model. We see that the different segments capture different regimes of the system, for example separating between class hours (mostly localized interactions) and break times (more dense connections across partitions). Figure from the original paper.

This model has three control parameters: a distribution of job benefits (encoding different aspects like income, working hours, *etc.*), a job density across a geographical area, and the total number of commuters in the region. Similarly to what happens in the gravity model, this model describes an interplay between the benefits and costs of mobility.

Schelling's model of segregation is an agent-based model that seeks to explain how individual preferences and perceptions of difference can collectively lead to segregation. The model's rules involve agents making local decisions about whether to move based on the density of neighbors matching their demographic category. A nonlinear transition occurs at a mild parameter value controlling the agents' tolerance to dissimilarities with their neighbors, resulting in global segregation patterns [29]. This model has rarely been compared quantitatively with data. The fine granularity of the registration data—allowing for the study of district-to-district relocations within large municipalities—that will be used and the possibility to cross-index this data with demographic and socio-economic annotations provides a good opportunity to validate this model.

These models are widely used in applied research. In the context of our investigation of internal migrations in Austria, we aim to validate these models in a data-driven fashion. This will be achieved by formulating these models as generative processes.

It is easy to present an example of this generative formulation in the case of the gravity law. The usual formulation of this model is the functional form:

$$I_{ij} = K \frac{(M_i M_j)^\alpha}{d_{ij}^\beta} \tag{1}$$

where I_{ij} is the magnitude of the flow of migrations between locations i and j, M_i and M_j are the respective population sizes and d_{ij} is the geographical distance separating them. The common approach in applied research is to take the natural logarithm of both sides of Eq. (1) and estimate the parameters K, α and β by finding the best fit of the resulting functional relation on the data. This procedure introduces several misspecifications such as the way zero values are treated or the biases created by the logarithmic transformation [30]. These problems can be overcome by modeling the count of migrations between two locations as a Poisson random variable, with expected value given by the functional form of the gravity law (a model of this kind was originally formulated in Ref. [31]):

$$P(I_{ij}|\lambda_{ij}) = \frac{e^{\lambda_{ij}} \lambda_{ij}^{I_{ij}}}{I_{ij}!} \text{ with } \lambda_{ij} = K \frac{(M_i M_j)^\alpha}{(d_{ij}+c)^\beta} \tag{2}$$

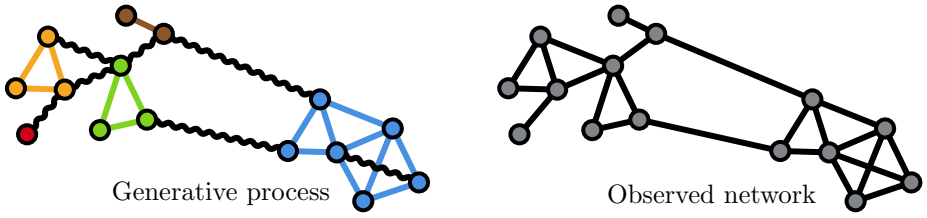

Figure 5: **Mixed effective-mechanistic models**. In this class of models, the generative process (left) encodes edges due to the mechanistic model under study (the squiggly edges) and models the structure of the residual edges (colored). In the observed network (right), the identity of the edges is erased. The inferential step consists of going backward and recovering the most likely generative process from the observed network.

Notice that in Eq. (2) we have introduced an additional parameter c. This parameter corrects for deviation at small distances and mobility within the same location (*i.e.* people changing their main residence within the same municipality).

This formulation of the mechanistic models in terms of generative distributions allows us to fully characterize the parameters in play by sampling from the posterior distribution, and not only recover point estimates. Moreover, it allows for testing the effect of modifications of the parameters of the models, effectively allowing for the simulation of external modifications such as the implementation of policies or the presence of other exogenous phenomena influencing mobility.

c. Mixed effective-mechanistic models for internal migrations in Austria

The effective and mechanistic modeling approaches presented above are typically pursued as separate and disconnected approaches. In this project, we will also propose to develop novel mixed effective-mechanistic models. This new class of model allows at the same time for a principled data-driven validation of the mechanistic models and for a meaningful and interpretable description of the patterns that they do not explain.

Following the inferential approach that we described above this class of mixed models will be defined through generative models encoding explicitly the mechanistic and effective contributions (Fig. 5). We will combine the effective methods considered previously and the mechanistic ones into a mixture model where the probability of the placement of an edge is influenced by a contribution of one of the mechanisms considered above together with structured "residuals" that are parametrized via the effective models. With this method, if the data can be explained perfectly with one of the aforementioned mechanisms, then the residual contribution will be empty. Otherwise—which is the expected scenario—we would be able to disentangle these underlying mechanisms from other processes that also affect migration patterns. These other processes will be captured by the effective part of the model.

Empirical motivation: localized communities and the effect of administrative boundaries. A meaningful example of the importance of this class of mixed models in the context of our work is found when looking at the communities in the network of relocations of Austrian residents in 2021. When fitting a hierarchical DC-SBM [32] we find two salient patterns in the community structure (Fig. 6). Firstly, we see that the communities are strongly localized in space, meaning that the people tend to move to municipalities close to their source location. Secondly, we see that there is a strong effect of the administrative boundaries (both at the level of federal states and of districts) on the community structure. While the first effect is correctly explained by the gravity law, the boundary effect—with municipalities geographically close but separated by an administrative boundary showing fewer relocations than would be expected—goes against it. In this case, a mixed model encoding the gravity law as its mechanistic

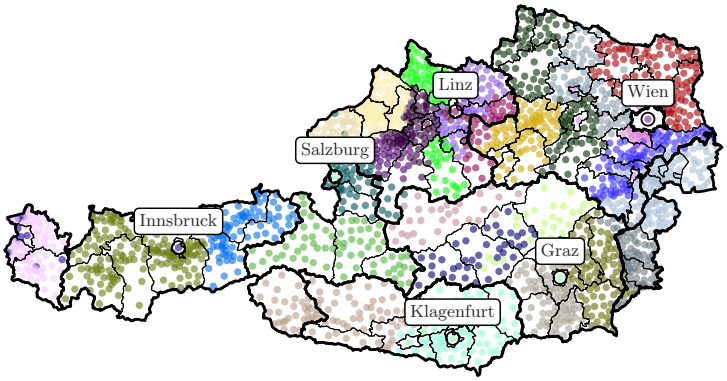

Figure 6: **Communities in the network of internal migrations in Austria in 2021.** The different municipalities are colored with unique colors corresponding to the community assignments (in layer $L = 1$) found by fitting a hierarchical DC-SBM [32].

part would be able to describe through its effective component this unexpected pattern as well as other effects not seen here and not captured by the gravity law.

Example: Gravity law and SBM. We present a simple example of how this class of mixed effective and mechanistic models can be implemented. In this case, we show the instance in which there is a cumulative relation between the two components; alternative types of relations can be formulated as well. We consider the observed network $\mathbf{A}$ as the outcome of a generative process where the weight of edges are (conditionally on the latent parameters) independent Poisson random variables:

$$P(\mathbf{A}|\{\theta\};\{M_i\},\{d_{ij}\}) = \prod_{i,j} \frac{e^{\lambda_{ij}}\lambda_{ij}^{A_{ij}}}{A_{ij}!} \tag{3}$$

The mechanistic-effective modeling is achieved by defining the expected weight of each edge as the sum of a mechanistic term encoding the gravity law and of an effective term given by the Poisson formulation of the DC-SBM:

$$\lambda_{ij} = \lambda_{ij}^{\mathrm{g}} + \lambda_{ij}^{\mathrm{s}} \text{ with: } \begin{cases} \lambda_{ij}^{\mathrm{g}} = K\frac{(M_iM_j)^{\alpha}}{(d_{ij}+c)^{\beta}} & \text{(mechanistic)} \\ \lambda_{ij}^{\mathrm{s}} = \omega_{b_ib_j}k_ik_j & \text{(effective)} \end{cases} \tag{4}$$

The Poisson formulation allows us to write the observed graph as the sum of two independent graphs, one completely described by the gravity law and the other completely described by the SBM:

$$A_{ij} = G_{ij} + U_{ij} \tag{5}$$

with:

$$P(\mathbf{G},\mathbf{U}) = \prod_{i,j} \frac{e^{\lambda_{ij}^{\mathrm{g}}}\lambda_{ij}^{\mathrm{g},G_{ij}}}{G_{ij}!} \prod_{i,j} \frac{e^{\lambda_{ij}^{\mathrm{s}}}\lambda_{ij}^{\mathrm{s},U_{ij}}}{U_{ij}!} \tag{6}$$

Intuitively, the observed network $\mathbf{A}$ is a simple superposition of the two networks $\mathbf{G}$ and $\mathbf{U}$ where we sum the weight of the parallel edge (conserving the directionality) and erase the identity of the edges (*i.e.* "forgetting" whether they are due to the gravity law or to the SBM). Given this model, the inferential task will deal with the posterior distribution:

$$P(K,\alpha,\beta,c,\mathbf{b},\mathbf{e},\mathbf{G},\mathbf{U}|\mathbf{A};\{M_i\},\{d_{ij}\}) \tag{7}$$

The effective part of the model will be able to describe meaningful structure in the residues of the mechanistic part, for example identifying long-range relocation preferences not explained by the gravity law.

IV. Workplan

The proposed Ph.D. project will contribute to the research project "Multiscale network modeling of migration flows in Austria", funded by WWTF. The team involved in this project comprises the two supervisors—Prof. Peixoto and Prof. Karsai—, Prof. Mathias Czaika from the University for Continuing Education Krems—researcher in social sciences and expert in migration phenomena—, and postdoctoral researcher Dr. Martina Contisciani. The candidate will work under the supervision of the two supervisors and in close collaboration with the postdoctoral researcher.

In the following, we present a provisional timeline of the project and a preliminary list of the potential risks related to the project.

a. Provisional timeline

The timeline is articulated in 6 semesters, covering the three years between the comprehensive exam and the thesis discussion. We indicate in the timeline only the milestones and activities in the Ph.D. itinerary that are specifically relevant to the project accomplishment; other relevant activities (*e.g.* teaching activities, outreach, *etc.*) are omitted.

1st semester	• Review of the literature on generative models for temporal graphs • Development of generative models for temporal graphs with community structure and multiscale temporal evolution
2nd semester	• Validation of the generative models on existing datasets of temporal graphs • Inference of the generative models on the aggregate data on migration flows
3rd semester	• Review of the literature of mechanistic models • Fitting of the mechanistic models on the aggregate data
4th semester	• Inference of the generative models on the full granularity dataset • Fitting of the models of mobility and segregation on the full granularity dataset • Development of the mixed generative-mechanistic model
5th and 6th semester	• Inference of the mixture models • Integration of the models with demographic annotations on the dataset • Final analysis of the results • Thesis writing

b. Feasibility, risk management, and mitigation strategies

This project is ambitious in its objectives, and with this comes risks commensurate with the potential rewards. Here, we list the main risks we identify and how they could be mitigated.

- **Intractability of the general modeling premise.** The data-driven analysis will be carried out using Bayesian inference. Extending the generative model framework with multiple dimensions and an effective-mechanistic modeling mixture could make the inference problem more complex. However, we are confident that this will not pose a significant obstacle. Previous research has demonstrated the feasibility of extending this modeling approach based on the stochastic block model to higher dimensions, including multilayer and temporal networks [16, 18], without compromising performance. Moreover, the Bayesian approach we intend to follow would automatically determine the model's complexity from the data. A simpler representation will be favored if the data does not require a more complex model. However, if our initial attempts at parameterizing the models underperform or turn out to be intractable, we can easily explore dimensionality reduction techniques or identify additional constraints to improve expressiveness. *The impact of this risk is high, but its probability is moderate to low.*

- **Algorithmic complexity.** The inference of the generative models and the mixture model will require using existing inference algorithms (based on MCMC, expectation-maximization, *etc.*) and their adaptation to the specific case under study. The computational feasibility of these tasks will depend on the complexity of the models used and on how many data layers will be integrated into them. In this kind of inferential task, there will be a trade-off between the model's expressive power (*i.e.* how complex are the features that the model can reproduce) and the computational complexity of the inferential task, and thus, eventually, it will be possible to simplify the modeling premises to overcome algorithmic complexity barriers. The computational aspects of this project will be carried out relying on existing tools [33] that already employ efficient algorithmic implementations [34] and allow for the parallelization of the more computationally intensive tasks. *The impact of this risk is moderate to low, and its probability is moderate.*
- **Slow progress and unexpected technical barriers.** The proposal is ambitious. It contains two different methodological contributions: the development of new generative models for graphs and the formulation of a mixture model combining them with mechanistic models of mobility. It may be possible that some of the activities prove more cumbersome than foreseen or that they provide rich results for the study of internal migrations in Austria that end up demanding more time for their analysis than initially planned. The distinct but complementary nature of the two methodological contributions is intended precisely to give enough flexibility to deal with such situations to adapt the progress as necessary. Moreover, the supervisors' and other researchers' experience will help make timely progress very likely. *The impact of this risk is low, and its probability is low.*

References

[1] D. G. Papademetriou and P. L. Martin, *The unsettled relationship: Labor migration and economic development*, 33 (Greenwood Publishing Group, 1991).

[2] M. Czaika and C. Reinprecht, in *Introduction to migration studies: An interactive guide to the literatures on migration and diversity* (Springer International Publishing Cham, 2022) pp. 49–82.

[3] P. Scholten, A. Pisarevskaya, and N. Levy, in *Introduction to Migration Studies: An Interactive Guide to the Literatures on Migration and Diversity* (Springer, 2022) pp. 3–24.

[4] J. Klugman, Overcoming Barriers: Human Mobility and Development (October 5, 2009). UNDP-HDRO Human Development Reports (2009).

[5] E. Tranos, M. Gheasi, and P. Nijkamp, Environment and Planning B: Planning and Design **42**, 4 (2015).

[6] V. Danchev and M. A. Porter, Social Networks **53**, 4 (2018).

[7] Y. Leo, E. Fleury, J. I. Alvarez-Hamelin, C. Sarraute, and M. Karsai, Journal of The Royal Society Interface **13**, 20160598 (2016).

[8] J. Mooyaart, M. Danko, and M. Boissonneault, Netherlands Interdisciplinary Demographic Institute (2021).

[9] G. Aristotelous, P. W. Smith, and J. Bijak, QuantMig Project Deliverable D6.3. (2022).

[10] P. Holme and J. Saramäki, Physics reports **519**, 97 (2012).

[11] T. P. Peixoto, Physical Review X **12**, 011004 (2022).

[12] L. Peel, T. P. Peixoto, and M. De Domenico, Nature Communications **13**, 6794 (2022).

[13] T. P. Peixoto, "The netzschleuder network catalogue and repository," (2023).

[14] N. Perra, B. Gonçalves, R. Pastor-Satorras, and A. Vespignani, Scientific reports **2**, 469 (2012).

[15] P. Holme, PLoS computational biology **9**, e1003142 (2013).

[16] T. P. Peixoto and M. Rosvall, Nature communications **8**, 582 (2017).

[17] X. Zhang, C. Moore, and M. E. Newman, The European Physical Journal B **90**, 1 (2017).

[18] T. P. Peixoto, Physical Review E **92**, 042807 (2015).

[19] L. Peel and A. Clauset, in *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 29 (2015).

[20] T. P. Peixoto and M. Rosvall, in *Temporal network theory* (Springer, 2023) pp. 65–82.

[21] T. P. Peixoto, in *Advances in network clustering and blockmodeling* (Wiley Online Library, 2019) pp. 289–332.

[22] T. P. Peixoto, Physical Review E **95**, 012317 (2017).

[23] T. P. Peixoto and L. Gauvin, Scientific reports **8**, 15511 (2018).

[24] J. Fournet and A. Barrat, PloS one **9**, e107878 (2014).

[25] J. J. Lewer and H. Van den Berg, Economics letters **99**, 164 (2008).

[26] F. Simini, M. C. González, A. Maritan, and A.-L. Barabási, Nature **484**, 96 (2012).

[27] T. C. Schelling *et al.*, Journal of mathematical sociology **1**, 143 (1971).

[28] R. Prieto Curiel, L. Pappalardo, L. Gabrielli, and S. R. Bishop, PloS one **13**, e0199892 (2018).

[29] L. Dall'Asta, C. Castellano, and M. Marsili, Journal of Statistical Mechanics: Theory and Experiment **2008**, L07002 (2008).

[30] M. Burger, F. Van Oort, and G.-J. Linders, Spatial economic analysis **4**, 167 (2009).

[31] R. Flowerdew and M. Aitkin, Journal of regional science **22**, 191 (1982).

[32] T. P. Peixoto, Physical Review X **4**, 011047 (2014).

[33] T. P. Peixoto, "The graph-tool python library," (2014).

[34] T. P. Peixoto, Physical Review E **89**, 012804 (2014).