The Virtue of High Performance Computing for the Statistical Analysis of Social Networks

Lars Leszczensky and Sebastian Pink

Mannheim Centre for European Social Research, University of Mannheim

We illustrate the potential of the bwHPC for sociological research. After briefly explaining why modern sociological research requires more computational time, we outline a recently developed computationally intense statistical model for the analysis of relational data. We close by summarizing an empirical study in which we applied this model to answer a substantive question and greatly benefited from the bwHPC with regard to computation time.

1 Introduction

Most sociological statistical analyses are conveniently estimable on single-core office computers. Examples for typical analyses are linear or logistic regression models that are estimated based on a survey of not more than 5,000 individuals, which take only a few minutes to run. However, a growing number of sociologists uses simulation techniques to answer substantive questions and this tremendously increases the computational resources needed. The reason is that simulation incorporates randomness and, consequentially, one run of a model is not sufficient. Instead, multiple runs of the same model specification are carried out and averaged afterwards to condense information. For such computationally intense simulations the distributed bwHPC cluster resources are of great use as they save lots of computation time.

Within a project funded by the German Research Foundation, we are interested in friendship formation among around 2,000 adolescents, which we observed in schools in North Rhine-Westphalia. We surveyed multiple grade-level networks in which we collected data on friendship nominations of all students within an entire grade (i.e., students of different classrooms within the same academic year could nominate one another). Furthermore, we repeated these surveys several times after roughly nine months each, so that we are able to follow the evolution of these friendship networks over time [4].

An appropriate statistical model to analyze such network panel data is the so-called stochastic actor-oriented model for the co-evolution of networks and behavior (SAOM) [7]. In essence, this model traces the evolution of the friendship nominations over time by simulating individual friendship-making decisions of the students of one network according to theoretically important elements of this decision (e.g., friends having the same sex as oneself or friends of friends becoming friends).

On a conventional office computer these models take about four hours to run. However, the number of estimated models equals the number of the individual networks that were surveyed, in our case typically 13. This makes this kind of statistical analysis suitable for parallel computation. We distribute the estimation of each SAOM to one core of one bwHPC node. This means a thirteenfold increase, or in other words, we are able to retrieve the results after four hours instead of 52 hours. Importantly, for an scientific article many model specifications have to be tested so that this computational advantage is even further multiplied by the amount of these different specifications. After the individual results of the SAOM are retrieved, they are averaged over by means of meta-analysis [1, 2].

2 Stochastic Actor-Oriented Models for the Co-Evolution of Networks and Behavior

SAOM can be regarded as empirically calibrated agent-based simulation models [8]. The process of network evolution leading from an initial state of an empirically observed network to subsequent network states observed at later points in time is modeled as the outcome of a stochastic process that is based on a sequence of actors' decisions. Each actor repeatedly decides myopically whether to change (i.e., create or dissolve) one of his outgoing ties or to retain the status quo. The choice of one particular network state x for one actor i is determined by his so-called evaluation function,

$$f_i(\beta, x) = \sum_k \beta_k s_{ki}(x). \tag{1}$$

The evaluation function is a linear combination of k effects $s_{ki}(x)$ (e.g., structural, actor, and dyadic effects) and their respective statistical parameters β_k [7, p. 47]. The set of statistical parameters β is estimated by means of simulation. Based on raw estimates, parameters are subsequently iteratively adjusted [6]. After a set of statistical parameters has been identified and fixed, their covariance is estimated by simulating a rather large number of networks [5]. As a result, a SAOM for a single network yields a vector of parameter estimates β as well as a $k \times k$ covariance matrix. Afterwards, in a meta-analysis [1] the input data are a $n \times k$ matrix $\hat{\theta}$ of k parameter estimates for n networks as well as a set S of n $k \times k$ matrices representing the within-study covariance matrices of the parameter estimates.

3 Exemplary Research

We outline the application of such SAOM to answer a substantive research question. The statistical models were run on the bwHPC, more precisely on 13 processors of one node. This research effort, which was published the journal Social Networks [3], sheds light on differences in ethnic segregation with regard to the structural embeddedness of students in classrooms and the respective grades.

It is well-established that adolescents' school-based friendship networks tend to be segregated along ethnic lines. But few studies have examined whether variation in network boundaries affects the degree of ethnic friendship segregation. We argue that ethnic homophily (i.e., those with the same ethnicity, e.g., Turkish, Polish, or German, befriend each other) is more pronounced in grade-level than in classroom-level networks. We consider classrooms to be low-cost situations for students to initiate and maintain friendship ties, because access to and daily interaction with classmates are very easy. Students not only attend daily lessons together but are also exposed to the same obligations, have to tackle the same tasks (e.g., do the same homework), and share experiences related to class trips or other joint activities. Grade-level networks, by contrast, constitute a high-cost situation for the formation of friendships within school. Students in the same grade, who do not share a classroom, have fewer opportunities to meet, and share fewer common experiences than those who are bound together by a joint classroom. Within school they can meet during breaks, but while classmates can be contacted quite easily before and after lessons, contact with schoolmates in other classrooms requires more of an active effort (e.g., walking to other classrooms, waiting for students, or encountering other students who one barely knows). As a consequence, a student may only invest in a relatively costly cross-classroom friendship if he or she perceives this friendship to be particularly beneficial. Ethnic homophily may thus result in a higher percentage of same-ethnic friendships at the grade level than at the classroom level.

We empirically tested this hypothesis using two repeated observations of 13 grade-level networks. In line with our theoretical argument, the tendency to form same-ethnic friendships was indeed stronger at the grade level, which translated into stronger ethnic segregation in friendship networks at the grade level than at the classroom level. Technically, four preferences are of interest: creating same-classroom-same-ethnicity ties, same-classroom-different-ethnicity ties, different-classroom-same-ethnicity ties, and different-classroom-different-ethnicity ties. We identified the joint contribution of the respective effects by inserting their estimated mean parameters into the evaluation function. The resulting contributions to the evaluation function are 0.07 for same-classroom-same-ethnicity ties, 0.0 for same-classroom-different-ethnicity ties, -0.46 for different-classroom-same-ethnicity ties, and -0.82 for different-classroom-different-ethnicity ties. Accordingly, the difference in preferences for same-ethnic friendships compared to interethnic friendships at the grade level was, at 0.36 (-0.46-(-0.82)), more pronounced than the respective preferential difference at the classroom level at 0.07 (0.07-0.0). This means that having the same ethnic background was more important for friendship selection if two students were not in the same classroom. This finding confirmed our hypothesis of a higher ethnic segregation of grade-level friendships.

Acknowledgements

This research was supported by a grant from the German Research Foundation (DFG/KA 1602/6-1; 6-2).

References

[1] An, W. “Multilevel meta network analysis with application to studying network dynamics of network interventions.” Social Networks (2015): 48-56.

[2] Gasparrini, A., Armstrong, B., and M. G. Kenward. “Multivariate Meta-Analysis for Non-Linear and other Multi-Parameter Associations.” Statistics in Medicine, 31, (2012): 3821-3839.

[3] Leszczensky, L. and S. Pink. “Ethnic Segregation of Friendship Networks in School: Testing a Rational-Choice Argument of Differences in Ethnic Homophily Between Classroom- and Grade-Level Networks.” Social Networks, 42, (2015): 18–26.

[4] Leszczensky, L., Pink, S. and F. Kalter. “Friendship and Identity in School: Field Report on

Wave 1, Wave 2, and Wave 3 (Technical Report).” Working Paper 161. Mannheim, Germany: Mannheimer Zentrum für Europäische Sozialforschung. (2015).

[5] Ripley, R. M., Snijders, T. A. B., Boda, Z., Vörös, A., and P. Preciado, P. “Manual for RSIENA (Version September 21, 2015).” University of Oxford, Nuffield College, Oxford, UK. (2015).

[6] Snijders, T. A. B. “The Statistical Evaluation of Social Network Dynamics.” Sociological Methodology, 31, (2001): 361-395.

[7] Snijders, T. A. B., van de Bunt, G. G., and C. E. G. Steglich. “Introduction to Stochastic Actor-Based Models for Network Dynamics.” Social Networks, 32, (2010): 44-60.

[8] Snijders, T. A., and C. E. G. Steglich. “Representing Micro–Macro Linkages by Actor-Based Dynamic Network Models.” Sociological Methods & Research, 44, (2015): 222-271.