# Who was Where, When?

## Spatiotemporal Analysis of Researcher Mobility in Nuclear Science

Miray Kas [#1], Kathleen M. Carley [#2], L. Richard Carley [#1]

[#1] *Electrical and Computer Engineering, Carnegie Mellon University*
[#2] *Institute for Software Research, Carnegie Mellon University*
*Pittsburgh, PA USA*

mkas@andrew.cmu.edu, kathleen.carley@cs.cmu.edu, carley@ece.cmu.edu

*Abstract*— **With the emergence and growth of digital libraries, information on collaboration among researchers has become very precise in terms of time and space. Digital libraries are increasingly used for extraction of co-authorship networks where two authors are connected by an edge if they have co-authored at least one paper. Co-authorship networks that span multiple years of research publications are 'spatially-embedded dynamic networks' because each publication has supplementary information about its publishing date and the affiliation(s) of its author(s). Spatiotemporal analysis on such dynamic co-authorship networks enables us to understand the '*mobility patterns*' of researchers. For instance, for each author, it is possible to identify if he/she has relocated to another institution or to another city/country by keeping track of the affiliations listed for him/her in the papers he/she has written over time. Mobility patterns provide answers to many interesting research questions such as which countries are the key influencers for other countries, how knowledge disseminates at the global level through the relocation of researchers, how collaboration patterns change over time and space as well as many others. In this paper, we focus on these research questions and perform spatiotemporal analysis on 20-year publication data in the field of nuclear physics, collected from arXiv pre-print library. Our results suggest that the USA and Germany are the countries that are most involved in nuclear physics researcher exchanges, and additional patterns of historical, geographical proximity, and language-based effects are observed in the mobility patterns of nuclear researchers across countries.**

## I. Introduction

Recent research on social networks is increasingly interested in analyzing large-scale social networks that have nodes on the order of thousands, millions. Scientific networks that are extracted from research publications are one kind of such large(r)-scale social networks that are of interest to numerous research fields such as social network analysis, bibliometrics, data mining, information retrieval, statistical physics and many others. One reason as to why scientific networks are extensively studied is that they provide tangible information about the interaction and collaboration patterns among social actors (e.g. authors) through the documentation of the papers they publish together.

Many types of social relationships including co-authorship are inevitably intertwined with time and space, both of which are important dimensions shaping who people interact with and build acquaintanceship/collaborations.

The growing availability and the ease of extracting detailed information on time and space dimensions of social relations enables researchers to build spatially-embedded dynamic networks and perform thorough spatiotemporal analysis on them. Spatiotemporal analysis is a multi-dimensional analysis method that can provide answers to multiple questions simultaneously, which would have been impossible if one of these two critical dimensions (e.g. time or space) has been neglected. For instance, using spatiotemporal analysis techniques enables researchers to understand and predict answers for generic questions such as 'Which locations will be the best/worst places to be at when?', 'Will there a particular location people will gravitate towards?', or 'Are there similarities in movement patterns of social actors?', all of which contribute to shaping the future of the community.

In this paper, we perform detailed spatiotemporal analysis on the mobility patterns of the authors extracted from 20-year publication data in the fields of experimental and theoretical nuclear physics, collected from arXiv pre-print library [1]. Our goal is to examine this data in detail by constructing transition matrices and trails per author and find answers to various questions such as which countries are the key influencers for others, how this is reflected on researcher mobility, how knowledge disseminates at the global level through the relocation of researchers, and what are the dominating patterns of the researchers' mobility.

## II. Related Work

While spatial networks are widely used in economics and economic geography, it is relatively recent for large-scale social network data that has thousands, millions of nodes. In this section, we briefly point out to a few studies in this area. [2] focuses on the problem of spatiotemporal community detection using a mobile phone dataset and proposes a modularity function adapted to spatial networks. [3] considers interactions and acquaintanceship of people on German collegiate social networking site called StudiVZ, focusing on how interactions are affected by geographical separations. [4] investigates the problem of discovering cohesive subgroups that are temporally and spatially close and have homogenous behaviour in time and space. [5] is a publication from Facebook which attempts to improve predictions on geographical locations of the users using their social and spatial proximity. [6] discusses representing network data that has time and space information in the form of time-aggregated graphs to reduce storage cost and increase scalability.

## III. Dataset and Software Technology

**Dataset:** In this paper, we use a dataset we have compiled using experimental and theoretical nuclear physics papers that are publicly available on the Cornell University pre-print archive (*i.e.*, arXiv website) [1]. In our dataset, there are 21,080 papers in nuclear physics. Our dataset spans nuclear physics publications that were added to the arXiv website between 1992 and 2010. Each paper has a unique identifier whose first four digits represent the year and month of publication, making it possible to track temporal dimension using paper IDs only.

In addition to providing the text and identifier information for the paper, the arXiv website facilitates bulk download of metadata about the paper in the form of XML files. In these metadata XML files, we have author and affiliation information which for each paper. However, there are three major issues that need to be addressed before this data can be used for analysis.

First, affiliation information is not a mandatory field; hence not all records include this information. We have gone through additional manual processing to make sure that we capture affiliation information for as many entries as possible.

Second, the same organization might be listed under different forms which would require uniformization to prevent emergence of redundant nodes in the network. For instance, Carnegie Mellon University might be listed as CMU, Carnegie Mellon, and Carnegie Mellon University in different papers. We have processed the affiliation names using a thorough organization thesaurus to convert different representations of the same organization to a uniform organization and performed manual processing to fix the entries that are not captured by the thesaurus.

The third issue is similar to the second issue we have: disambiguation of the author names and similar processing is performed to cleanup author entries as well. In the clean form of our dataset, we have 16,404 authors.

**Software:** We perform spatiotemporal analysis using ORA-GIS and Loom platforms that facilitate analyzing the movements of social actors through a set of named locations (*i.e.*, a discrete state space) over time [6].

## IV. Methods and Results

In this section, we discuss the spatiotemporal analysis methods we have used and present the corresponding results. We primarily focus on presenting results at the country level.

### A. ***Method-1:*** *Aggregate Transition Matrix*

As a first step in our analysis, we construct an 'aggregate transition matrix' across countries. In this matrix, rows and columns represent countries and there is a link from *Country-A* to *Country-B* if a social agent has travelled directly from *Country-A* to *Country-B*. We start with constructing an Author x Affiliation matrix. We obtain the relocation information of researchers (either temporary or permanent) through their affiliations listed in the papers they have co-authored. Then, using the (Affiliation x Country) matrix that represents which organization is in which country, we obtain (Author x Country) matrix by multiplying (Author x Affiliation) and (Affiliation x Country) matrices. This gives us spatial dimension of our network where the countries constitute the discrete space set we interested in performing analysis on.

The over-time, dynamic dimension is achieved by generating multiple (Author x Country) matrices each of which model a subset of publications that are published in a certain time range (month/year). After obtaining an (Author x Country) matrix for each time interval (let's say a year), we are able to construct the 'transition matrix'. Consider the following scenario. In 1999, *Author-A* was affiliated with an organization in *Country-A*. And, in 2000, the same author, *Author-A*, appears as affiliated with another organization in *Country-B*. Then, in the transition matrix, we insert a link from *Country-A* to *Country-B* because a social agent (e.g. *Author-A*) has travelled directly from *Country-A* to *Country-B.*

In our results, we call the resultant transition matrix to be an 'aggregate' matrix as it consolidates the spatiotemporal transition information of all authors in our dataset.

### B. ***Results-1:*** *Transition of Researchers across Countries*

Figure 1 presents a visualization of aggregate transition matrix and illustrates which countries are actively involved in researcher exchanges.

The links drawn in the network presented in Figure 1 are directed, and as we discussed earlier, there exists a link from country *A* to *B* if a researcher from country *A* has relocated to country *B*. Let's walk through an example we have observed in our dataset. For instance, an author, *researcher-X* was affiliated with Massachusetts Institute of Technology, Cambridge, MA, USA, and at that point we consider *researcher-X* to be from the USA, regardless of ethnic origin. Then, *researcher-X* has relocated to Niels Bohr Institute, University of Copenhagen, Denmark. With this change in the network, we update the affiliation of *researcher-X* at organization, city, state, and country levels. Such a transition from the USA to Denmark would insert a link in Figure 1 directed from the USA to Denmark.

In Figure 1, in addition to the directions of the links, the size and colors of the nodes have meanings as well. We have used a spectrum of colors (e.g. red to brown to green to turquoise to blue) to quantify the number of people that leave a country to join another organization located in a different country. The closer the color of a country node is to blue, the more people go abroad from that country. On the other hand, we use node sizes to quantify the number of researchers a country accepts from institutions abroad. The country nodes that are drawn larger in Figure 1 represent the countries that receive more researchers from foreign organizations.

One thing to note Figure 1 is that the countries that are the highest senders are also the highest receivers (USA and Germany). This is primarily due to many researchers going

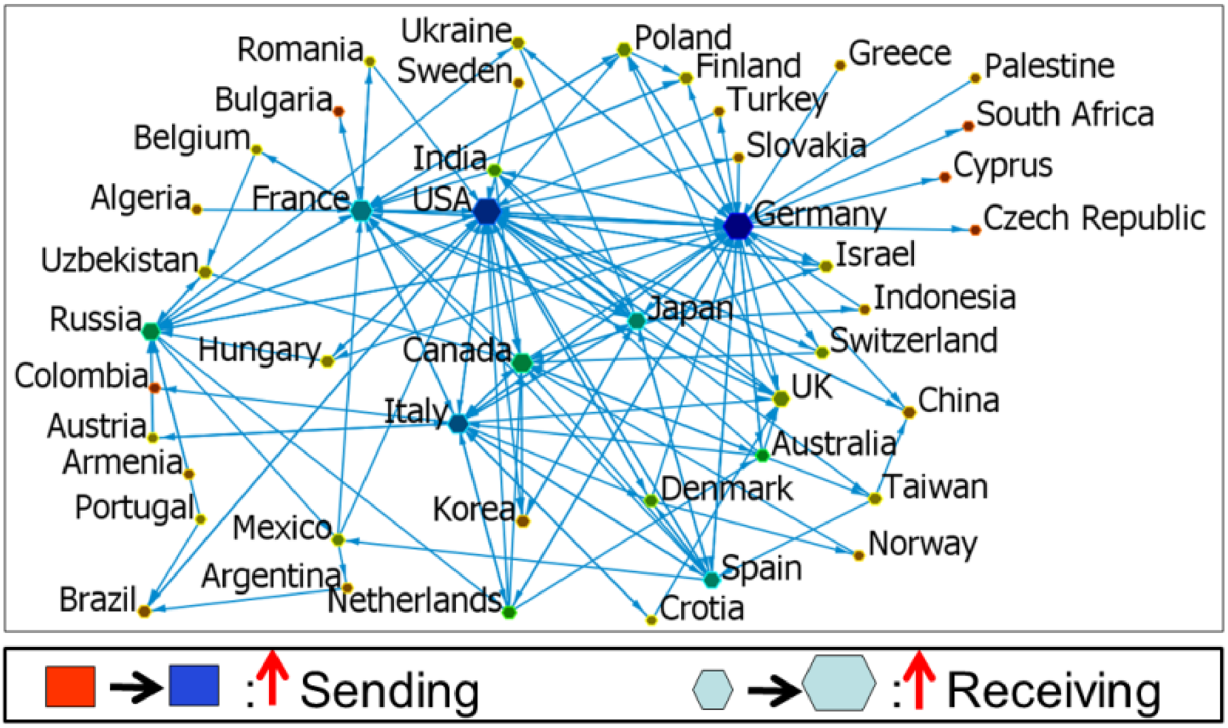

**Figure 1 - Spatiotemporal Researcher Mobility Graph (At Country Level)**

back to their own countries, which we discuss more in our trail analysis results. Following the USA and Germany, there are other countries that are visible as key players in researchers' mobility such as Italy, France, Japan, Canada, Russia, and Spain.

There are other interesting results that can be inferred from Figure 1. For instance, speaking the same/similar language makes researcher exchange easier; it is remarkable how clearly this pattern comes out in Figure 1. Brazil, Argentina, Mexico, Portugal, and Spain exchange researchers very frequently, with geographical proximity being an additional factor for further grouping within this subset of countries. Another example to the impact of geographical proximity is the connection between two European countries such as Croatia and Italy.

From Figure 1, it is also possible to note political, historical relations as another factor contributing to researcher exchanges. For example, Algeria was a French colony for a long period, and a significant number of Algerians were able to get French citizenship. Algeria still has close intellectual/cultural relations with France, and it is not surprising to see an arrow directed from Algeria to France as the only transition link emanating from Algeria.

In addition, looking at the overall structure of the network, one might notice a core-periphery network kind of structure where there is a set of core nodes that are well-connected with one another, and also with the periphery. In contrast, peripheral nodes are connected to the core, but not connected to one another. In this network, the core nodes are mostly the major players that are long known to have nuclear capabilities (e.g. USA, France, and Japan). The nodes that are towards the peripheral side of the network are mostly emerging or small(er) countries, which have fewer connections to other countries. As one final note, one might notice Palestine as a peripheral country. Palestine appears in this graph because of a Catholic University in the West Bank that is actively publishing papers. Palestine is one special case for which the publication patterns do not tell us much about the capabilities of the country itself.

### C. ***Method-2:*** *Trails Analysis*

The next spatiotemporal analysis method we apply is '*trails analysis*', which tracks movements of individual actors over time, and outputs 'who was where, when?' information in the form of traces (trails) per actor.

Trails are the paths the social agents move through within a network and they provide powerful representations of spatiotemporal network data. The visual representation of trails analysis involves drawing a big table where the column represent the discrete space set –locations (e.g. countries)– and the rows represent the discrete time set. Each social agent –in our case, each author– has a sequence of locations he/she was present at different points in time and the path he/she followed over time is printed as a path, as a *trail*.

### D. ***Results-2:*** *Dominant Patterns in Researcher Mobility*

We next present a visually comprehensible subset of information we obtained via trails analysis, and only depict a representative set of authors' trajectories to avoid making the figure (Figure 2) over complicated.

In Figure 2, each drawn line (differentiated by color) represents another actor. Columns represent countries that are most actively involved in researcher exchanges while rows represent the years during which the data is collected.

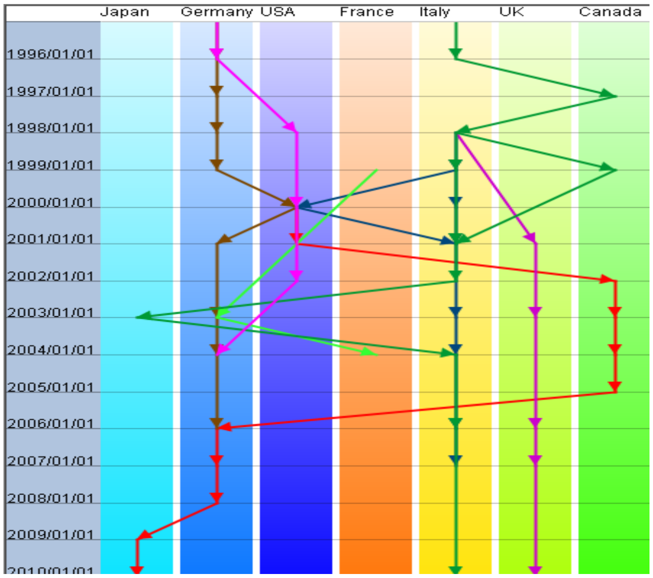

**Figure 2- Dominating Trail Patterns in Researcher Mobility**

Similar to the generation of aggregate transition matrix, the trail information of each author is again collected based on authors' affiliations. In addition to the authors that move from country to country, there are also many authors that tend to stay in the same country during their entire career. We preferred not to depict such geographically localized authors' patterns in our results. For anonymity purposes, we also do not give out authors' names; we instead refer to them in terms of the colors their corresponding trails are represented with.

For instance, the authors represented in pink and brown start their careers in Germany and then move to the USA, and finally they move back to their countries. This pattern (*i.e.*, returning to the origin country after visiting a major country) is actually one of the most dominating patterns that emerge from our dataset. This is primarily an effect of the physics research field. Authors returning to their origin country after working as post-doctoral researchers or visiting scholars at foreign institutions are very commonly observed. The authors represented in pink and brown are such authors who got their Ph.D.'s from Germany, and returned to Germany after in the USA for a while. If they were to start their research careers in the USA, they would not probably have prior publications with Germany based affiliations.

The red line and the dark green line represent authors whose main affiliations are non-academic. They are primarily involved in big collaborations or national laboratories. For non-academic, industrial authors, we have observed that mobility is something that significantly contributes to their productivity; increasing the number of papers they write.

## V. Conclusions and Future Work

In this paper, we perform spatiotemporal analysis on the mobility patterns of researchers that publish in the fields of theoretical and experimental nuclear physics. Our results highlight the USA and Germany, followed by France, Italy, Canada, and Japan as the key countries that are actively involved in researcher exchanges and influential for other countries. Examining the researcher mobility patterns at the country level, we find that spatial proximity or speaking similar languages are among the contributors to researcher exchanges among countries. When we consider the same information at the individual level, we notice that returning to the country of origin after working as a postdoctoral researcher or visiting scholar at a foreign institute is a common pattern among nuclear physicists.

There are a number of possible future research directions. For instance, the evolution patterns for the topologies of co-authorship/citation networks might vary significantly across different fields and this has been previously discussed in many papers (e.g. Newman [7]). However, spatiotemporal patterns are yet to be studied to that extent. One potential direction is to perform spatiotemporal analyses on publications from different research fields similar to the analyses presented in this paper and provide a detailed comparison of patterns observed while discussing the potential reasons the differences and/or similarities arise from. Another very interesting line of research is to attempt to find governing equation(s) that might be useful for predicting the mobility patterns of researchers over time and the overall social impact of having researchers relocating.

## VI. Acknowledgements

Financial support was provided by the Defense Threat Reduction Agency (DTRA) under grant number HDTRA11010102. The views and conclusions contained in this document are those of the author and should not be interpreted as representing the official policies, either expressed or implied, of the Defense Threat Reduction Agency (DTRA) or the U.S. government.

## References

[1] Cornell University. (2011, January) Cornell University Library (arXiv). [Online]. http://arxiv.org/

[2] P. Expert, T. S. Evans, V. D. Blondel, and Lambiotte R., "Uncovering space-independent communities in spatial networks," *PNAS*, pp. 1-6, May 2011.

[3] C. Lee, T. Scherngell, and M. J. Barber, "Investigating an online social network using spatial interaction models," *preprint submitted to Social Networks*, November 2010.

[4] J. S. Yoo and J. Hwang, "A Framework for Discovering Spatio-Temporal Cohesive Networks," in *Proceedings of the 12th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, 2008, pp. 1056--1061.

[5] S. Shekhar and D. Oliver, "Computational Modeling of Spatio-temporal Social Networks: A Time-Aggregated Graph Approach," *Specialist Meeting - Spatio-Temporal Constraints on Social Networks*, 2010.

[6] G. B. Davis, J. Olson, and K.M. Carley, "OraGIS and Loom: Spatial and temporal extensions to the ORA Analysis Platform," CASOS, Carnegie Mellon University, Pittsburgh, Technical Report CMU-ISR-08-121, 2008.

[7] M.E.J. Newman, "Coauthorship networks and patterns of scientific collaboration," *Proceedings of National Academy of Sciences*, vol. 101, pp. 5200--5205, April 2004.